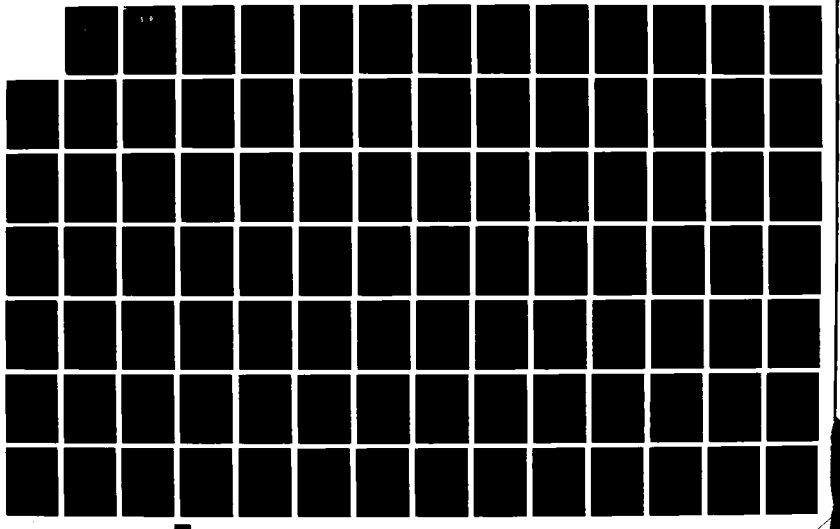
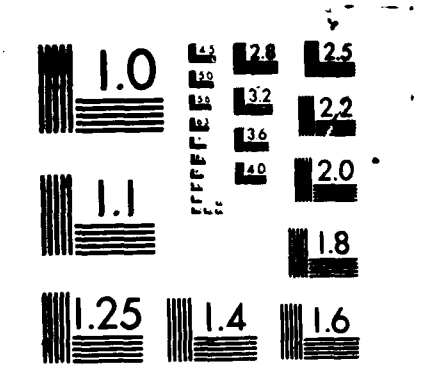
AD-A103 553 INTEGRATED PROCESSING IN PLANNING AND UNDERSTANDING(U) 1/3
YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE
L BIRNBAUM DEC 86 YALEU/CSD/RR-409 N00014-75-C-1111
UNCLASSIFIED F/G 12/9 NL





(2)

Integrated Processing in Planning and Understanding

Lawrence Birnbaum

YALEU/CSD/RR #489

December 1986

YALE UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 489	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Integrated Processing in Planning and Understanding	5. TYPE OF REPORT & PERIOD COVERED: Research Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Lawrence Birnbaum	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-1111, N00014-85-K-0108
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Yale University - Department of Computer Science, 10 Hillhouse Avenue, New Haven, CT 06520	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, VA 22209	12. REPORT DATE: December 1986
	13. NUMBER OF PAGES: 197
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems Program, Arlington, VA 22217	15. SECURITY CLASS. (of this report): unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Artificial Intelligence,
Language understanding,
Opportunistic planning,
Explanatory Inference
Freudian slips.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Programs that plan and understand must make many decisions about which paths of inquiry are likely to prove fruitful. In order to make such decisions rationally, and hence avoid the need for backtracking that inevitably results if they are made arbitrarily, relevant contextual information must be brought to bear. An integrated model of planning or understanding is one that attempts to take such contextual information into account as early as possible.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

An integrated model of understanding must take the understander's goals and hypotheses into account in making decisions about how to interpret an input. The relationship between syntax and semantics in language understanding is analyzed from such an integrated point of view. Next, the problems of lexical ambiguity and vagueness are addressed, previous attempts to solve these problems are analyzed, and their shortcomings are used to motivate requirements for a more complete solution. Finally, an integrated approach to inference in explanation-based understanding is presented.

An integrated model of planning must take the situation in which the planner finds itself into account early in the construction of plans. Such an integrated approach leads to a conception of planning in which goals are set, in part, on the basis of opportunities for their pursuit. This model of opportunistic planning is applied to the problem of response formation in arguments. The problems involved in recognizing opportunities are uncovered, and several possible solutions are presented. These solutions lead to a conception of unsatisfied goals as active mental entities. In this context, I analyze the planning architecture which is presupposed by Freud's intentional explanations for errors, particularly slips of the tongue, and show that it can be functionally justified on the grounds that it fulfills the requirements for opportunistic planning.

OFFICIAL DISTRIBUTION LIST

Defense Documentation Center Cameron Station Alexandria, Virginia 22314	12 copies
Office of Naval Research Information Systems Program Code 437 Arlington, Virginia 22217	2 copies
Dr. Judith Daly Advanced Research Projects Agency Cybernetics Technology Office 1400 Wilson Boulevard Arlington, Virginia 22209	3 copies
Office of Naval Research Branch Office - Boston 495 Summer Street Boston, Massachusetts 02210	1 copy
Office of Naval Research Branch Office - Chicago 536 South Clark Street Chicago, Illinois 60615	1 copy
Office of Naval Research Branch Office - Pasadena 1030 East Green Street Pasadena, California 91106	1 copy
Mr. Steven Wong New York Area Office 715 Broadway - 5tn Floor New York, New York 10003	1 copy
Naval Research Laboratory Technical Information Division Code 2627 Washington, D.C. 20375	6 copies
Dr. A.L. Slafkosky Commandant of the Marine Corps Code RD-1 Washington, D.C. 20380	1 copy
Office of Naval Research Code 455 Arlington, Virginia 22217	1 copy

Office of Naval Research — 1 copy
Code 458
Arlington, Virginia 22217

Naval Electronics Laboratory Center — 1 copy
Advanced Software Technology Division
Code 5200
San Diego, California 92152

Mr. E.H. Gleissner — 1 copy
Naval Ship Research and Development
Computation and Mathematics Department
Bethesda, Maryland 20084

Captain Grace M. Hopper, USNR — 1 copy
Naval Data Automation Command, Code 00H
Washington Navy Yard
Washington, D.C. 20374

Dr. Robert Engelmore — 2 copies
Advanced Research Project Agency
Information Processing Techniques
1400 Wilson Boulevard
Arlington, Virginia 22209

Professor Omar Wing — 1 copy
Columbia University in the City of New York
Department of Electrical Engineering and
Computer Science
New York, New York 10027

Office of Naval Research — 1 copy
Assistant Chief for Technology
Code 200
Arlington, Virginia 22217

Computer Systems Management, Inc. — 5 copies
1300 Wilson Boulevard, Suite 102
Arlington, Virginia 22209

Ms. Robin Dillard — 1 copy
Naval Ocean Systems Center
C2 Information Processing Branch (Code 8242)
271 Catalina Boulevard
San Diego, California 92152

Dr. William Woods — 1 copy
BBN
50 Moulton Street
Cambridge, MA 02138

Professor Van Dam — 1 copy
Dept. of Computer Science
Brown University
Providence, RI 02912

Professor Eugene Charniak — 1 copy
Dept. of Computer Science
Brown University
Providence, RI 02912

Professor Robert Wilensky — 1 copy
Univ. of California
Elec. Engr. and Computer Science
Berkeley, CA 94707

Professor Allen Newell — 1 copy
Dept. of Computer Science
Carnegie-Mellon University
Schenley Park
Pittsburgh, PA 15213

Professor David Waltz — 1 copy
Univ. of Ill at Urbana-Champaign
Coordinated Science Lab
Urbana, IL 61801

Professor Patrick Winston — 1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Marvin Minsky — 1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Negroponte — 1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Jerome Feldman — 1 copy
Univ. of Rochester
Dept. of Computer Science
Rochester, NY 14627

Dr. Nils Nilsson — 1 copy
Stanford Research Institute
Menlo Park, CA 94025

Dr. Alan Meyrowitz — 1 copy
Office of Naval Research
Code 437
800 N. Quincy Street
Arlington, VA 22217

LCOL Robert Simpson — 1 copy
IPTO-DARPA
1400 Wilson Blvd
Arlington, VA 22209

Dr. Edward Shortliffe — 1 copy
Stanford University
MYCIN Project TC-117
Stanford Univ. Medical Center
Stanford, CA 94305

Dr. Douglas Lenat — 1 copy
Stanford University
Computer Science Department
Stanford, CA 94305

Dr. M.C. Harrison — 1 copy
Courant Institute Mathematical Science
New York University
New York, NY 10012

Dr. Morgan — 1 copy
University of Pennsylvania
Dept. of Computer Science & Info. Sci.
Philadelphia, PA 19104

Mr. Fred M. Griffee — 1 copy
Technical Advisor C3 Division
Marine Corps Development
and Education Command
Quantico, VA 22134

Dr. Vince Sigilitto — 1 copy
Program Manager
AFOSR/NM
Bolling Airforce Base
Building 410
Washington, DC 20332

Integrated Processing in Planning and Understanding

Lawrence Albert Birnbaum
YALEU/CSD/RR #489

December 1986

This work was supported by the Defense Advanced Research Projects Agency (monitored by the Office of Naval Research under contracts N00014-75-C-1111 and N00014-85-K-0108) and the National Science Foundation (under grants IST7918463 and IST8017790).

ABSTRACT

INTEGRATED PROCESSING IN PLANNING AND UNDERSTANDING

Lawrence Albert Birnbaum

Yale University

1986

Programs that plan and understand must make many decisions about which paths of inquiry are likely to prove fruitful. In order to make such decisions rationally, and hence avoid the need for backtracking that inevitably results if they are made arbitrarily, relevant contextual information must be brought to bear. An *integrated* model of planning or understanding is one that attempts to take such contextual information into account as early as possible.

An integrated model of understanding must take the understander's goals and hypotheses into account in making decisions about how to interpret an input. The relationship between syntax and semantics in language understanding is analyzed from such an integrated point of view. Next, the problems of lexical ambiguity and vagueness are addressed, previous attempts to solve these problems are analyzed, and their shortcomings are used to motivate requirements for a more complete solution. Finally, an integrated approach to inference in explanation-based understanding is presented.

An integrated model of planning must take the situation in which the planner finds itself into account early in the construction of plans. Such an integrated approach leads to a conception of planning in which goals are set, in part, on the basis of opportunities for their pursuit. This model of *opportunistic* planning is applied to the problem of response formation in arguments. The problems involved in recognizing opportunities are uncovered, and several possible solutions are presented. These solutions lead to a conception of unsatisfied goals as active mental entities. In this context, I analyze the planning architecture which is presupposed by Freud's intentional explanations for errors, particularly slips of the tongue, and show that it can be functionally justified on the grounds that it fulfills the requirements for opportunistic planning.

INTEGRATED PROCESSING IN PLANNING AND UNDERSTANDING

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Lawrence Albert Birnbaum

December 1986

To my parents, Abe and Lore

ACKNOWLEDGMENTS

My advisor, Roger Schank, has been a mentor to me in every sense of the word. Roger first sparked my interest in AI when I was an undergraduate at Yale, and he has been an unflagging source of ideas, inspiration, and encouragement to me ever since. Above all, Roger taught me how to recognize interesting and important problems, and he encouraged me to pursue them. On a more personal note, I must thank him for his faith in me: I could not have finished this thesis without his friendship and support. Not the least of the debts I owe him is for the hospitality that he and his family showed me during a week in Paris when the final outline of this thesis was determined.

Drew McDermott has been the other major influence on my intellectual development. If I learned what counts as a question from Roger, I learned what counts as an answer from Drew. Through many helpful discussions, and by his example, he has challenged me to think more clearly about difficult issues.

Many thanks are also due to the other two members of my committee. Elliot Soloway provided moral support at crucial times. Robert Abelson helped to create, with Roger, an environment in which cognitive science could flourish, and his timely comments on my penultimate draft were enormously helpful.

Chris Riesbeck taught me AI programming, and he introduced me to the problems of language analysis. He also took the trouble to read a draft of this thesis and provide me with thoughtful comments.

I also want to thank all of my other teachers over the years at Yale, especially John Anderson, David Barstow, John Black, Guy Carden, Eugene Charniak, Dan Gusfield, Ned Irons, Wendy Lehnert, David Lichtenstein, and Alan Perlis.

To thank all of my fellow graduate students and colleagues would take far too much space. Conversations with Gregg Collins have given me more ideas than I can remember. The theory of opportunistic planning and Freudian slips, reported in chapter nine, was developed jointly with Gregg, and has appeared previously in Birnbaum and Collins (1984). The research on opportunistic planning in arguments, reported in chapter seven, grew out of joint work with Margot Flowers and Rod McGuire. I had several important discussions on

opportunistic planning with Natalie Dehn. I also want to thank Mark Burstein, Laurence Danlos, Ernie Davis, Kris Hammond, Ed Hovy, Larry Hunter, Alex Kass, Stan Letovsky, Steve Lytinen, Charlie Martin, Chris Owens, Jerry Samet, Colleen Seifert, and Mallory Selfridge for many helpful discussions. I had several useful discussions on language analysis with Mitch Marcus.

Friends make it all worthwhile. I especially thank Valerie Abbott, Charles Beasley, Mark Burstein, Gregg Collins, Heather Collins, Laurence Danlos, Ann Drinan, John Gipson, Shoshana Hardt, Larry Hunter, Stan Letovsky, David Lichtenstein, Rod McGuire, Jacob Mey, Debbie Samet, Jerry Samet, Diane Schank, Colleen Seifert, and Mallory Selfridge for their friendship. And special thanks to Amy Schwartz for her love and support.

Finally, I want to thank the agencies that provided financial support for the research reported here, the Defense Advanced Research Projects Agency (monitored by the Office of Naval Research under contracts N00014-75-C-1111 and N00014-85-K-0108) and the National Science Foundation (under grants IST7918463 and IST8017790).

TABLE OF CONTENTS

PART I

CHAPTER 1

RATIONAL DECISIONS AND INTEGRATED PROCESSING

1. Introduction

All intelligent programs must make decisions at many points in their processing. A language analyzer, for example, must decide what sorts of structures it is likely to find in an input utterance. A planner must decide which plan to try, from among all the possible ones, to achieve a given goal. And a plan understander must decide which inferences to draw in attempting to understand some input, given many possibilities. The need to make such complex decisions concerning which paths of inquiry are likely to prove fruitful, and are hence worthy of further pursuit, is a hallmark of problems in artificial intelligence.

The degree to which a program can make such decisions rationally is entirely dependent on the degree to which it has access to, and is able to take into account, the relevant information. Most of all, the program must be able to recognize the relevant information *as being* relevant. The problem is that such information often seems, at first, far removed from the choices being considered. It can thus be quite difficult to determine what factors need to be taken into account, and how. As a result, almost all current language analyzers, planners, and plan understanders make many or even most decisions about where to concentrate their efforts arbitrarily. They must make these decisions arbitrarily because, at the time and in the place in processing where the decision must be made, the information that might be relevant to making it rationally is either unavailable, or else no provision has been made for taking it into account.

The information that is relevant to deciding which plan to try to achieve a given goal, for example, may depend on features of the situation in which the planner currently finds itself. In particular, the resources that are available are a crucial determinent of whether or not a plan can be successfully executed. When planning to prepare dinner, for example, it is obviously worthwhile to consider the available food and utensils before plunging into details of menu or food preparation technique. Yet few, if any, theories of planning seem concerned with how such information can, or even *should*, be taken into account.

Similarly, the information that is relevant to deciding which inferences to draw in attempting to understand a given input may depend on the understander's own goals, and on its current hypotheses about the whole situation to which the input pertains. Whether or not an input can be explained in terms of those hypotheses is a crucial determinant of how well it can be understood, and how an hypothesis bears on the understander's own goals is a crucial determinent of whether it is worthwhile to keep it under consideration. When trying to understand another agent's behavior, for example, if a given action can be explained in terms of some goal which the understander already ascribes to that other agent, then there is little point in constructing hypothetical explanations in terms of some other goal. Yet, although many theories of understanding argue that such information *should* be taken into account, few address the question of how it actually *can* be taken into account.

To the extent that planners and understanders fail to take such relevant information into account when deciding which lines of inquiry to devote effort to, they must decide arbitrarily. The problem with making such decisions arbitrarily, of course, is that they usually turn out to be wrong. Thus, all models which make such decisions arbitrarily must employ one of two possible strategies to cope with the problems caused by numerous erroneous choices: Either they must depend on backtracking when choices prove to be erroneous or irrelevant, or they must ensure -- by fiat of the programmer -- that exactly and only the choices that will be relevant to the example at hand will in fact be available, in order to avoid having to make any decisions at all. In other words, without any theory of how a space of possible solutions *should* be searched, a model must either resort to exhaustive search, or else simply pretend that there is no search. The former leads to models which are at best inefficient, and at worst combinatorially intractable; the latter, to models which work only on one or, at the most, two examples.

For example, in the former category, most models of syntactic analysis -- e.g., ATNs and Prolog-based parsers (Thorne, Bratley, and Dewar, 1968; Bobrow and Fraser, 1969; Woods, 1970; Colmerauer, 1978; Pereira and Warren, 1980) -- spend an inordinate proportion of their time backtracking from decisions to pursue analyses that wouldn't make sense even if they succeeded, and that no human understander would bother pursuing. Many models of understanding get bogged down making enormous numbers of irrelevant inferences (e.g., Rieger, 1975), or pursuing inference paths towards explanations that don't apply in the current situation (as in, e.g., most work on plan understanding). Many models of planning, when confronted with choices, simply pursue all of them, and end up constructing many plans that cannot possibly work under the circumstances (e.g., Tate, 1977).

The latter category of programs which avoid search by the fiat of the programmer is in many ways even less principled. Planning models which take as their chief goal flexibility are much less flexible than hoped because many decisions are finessed, and, as a result, many plans that could otherwise be constructed are simply blocked (e.g., most NOAH-style hierarchical planners). Similarly, many language analyzers simply assume that much of the ambiguity inherent in language is just not present, for example by completely ignoring lexical ambiguity (Marcus, 1980). And, finally, many understanding models that seem to avoid search by bringing a great deal of knowledge to bear -- and are, indeed, even termed "integrated" -- are in fact purely bottom-up models that avoid search at the programmer's whim, and are therefore incapable of drawing the inferences necessary to understand even minor variations in their target input (e.g., Dyer, 1983).

It thus appears imperative for artificial intelligence programs to make decisions rationally. Such rationally taken decisions will not always be correct, of course, but they are far more likely to be correct than are arbitrary ones. The need to make rational decisions, and to thereby avoid the indiscriminate use of backtracking, has two immediate consequences for process models. First, rational decisions are *informed* decisions: The more information that is taken into account in a decision, the more rational that decision is likely to be, and the better the chances of making it correctly. Thus, a process which seeks to make rational decisions must be able to take into account as much potentially relevant information as possible -- that is, as much as seems cost-effective -- regardless of the source or type of that information. For an understanding program, this means that such contextual information as the goals of the understander in a situation, and its hypotheses about that situation, must be taken into account in making decisions about how to determine what an input means in that situation. For a planning program, this means that features of the world in which the planner finds itself must be taken into account in reaching decisions about which goals and plans to pursue and how to pursue them.

The second major requirement imposed on process models by the need to avoid backtracking stems from the fact that early decisions about which line of reasoning to pursue, if erroneous, can have a particularly devastating impact on overall performance. To the extent that later decisions depend on earlier ones, all of the effort expended subsequent to an incorrect decision may be irrelevant or even counterproductive. Furthermore, determining which decision was at fault may require re-examining many or most of the intervening decisions. Thus, in order to minimize as much as possible the number of erroneous decisions, and the backtracking that results, it is particularly important that *early* decisions about which lines of reasoning to pursue be taken as rationally as possible. In particular, then, decisions taken early

in the processes of planning and understanding must be based on as much relevant information as possible. Models which seek to meet the above criteria -- which take as much contextual information into account as possible, as early as possible -- are *integrated* models.

In order to construct an integrated model of some ability, several difficult questions must be addressed. First, of course, what sorts of decisions need to be made? Second, what kind of information is relevant to making those decisions? Third, how can that information be made available to the decision process? And fourth, how can the decision process be structured so as to take advantage of the available information? One way or another, any process model must address these questions. What makes integrated models different are the additional functional requirements imposed on the answers, stemming from the need to avoid arbitrary decisions and backtracking. In particular, in order to make decisions as rationally as possible, an integrated approach to planning or understanding attempts to construct models that can employ any knowledge that might be useful, from as wide a variety of sources as necessary.

There are, of course, alternative approaches to answering the above questions. Faced with the difficulty of identifying the applicability of, let alone actually employing, different sorts of information from a wide variety of sources in making decisions, it is tempting to hope that the information that is necessary to make decisions in a process is of limited scope and obvious relevance. It is this hope that lies at the root of non-integrated approaches to planning and understanding, which assume that the information that is relevant to making decisions will be, by and large, from a single source and of a single type. Such approaches are typically termed *modular*, and I will continue to follow this terminological convention here. However, I want to emphasize that the issue is not one of modularity versus non-modularity, but of how the modules are to be defined. In an integrated approach, the modules that make up a process are determined solely on functional grounds, in terms of the role they play in carrying out the goals of the process as a whole. Modules can and should take into account any information which seems relevant in making decisions about how to carry out their functions. In a non-integrated approach, on the other hand, modules are conceived as having access to only a single kind of information, from a single source. Indeed, the class of information with which a module is concerned -- a class which is typically circumscribed on *a priori* intuitive or descriptive grounds, rather than being functionally justified -- is taken to be *the* defining characteristic of that module.

Thus, in determining how to interpret a particular feature in some situation, modular theories of understanding will take into account the presence or absence of other features of the same type. What they will generally *not* take into account, however, is the role that *other* kinds

of knowledge and information will play in this process of interpretation. For example, modular theories of language analysis, typically concerned -- to the exclusion of all other problems in language analysis -- with determining the syntactic properties of input sentences, will attend primarily or exclusively to syntactic knowledge and information in attempting to perform that task. Semantic and pragmatic information is presumed to play a limited or non-existent role in determining sentence structure, despite the fact that such information is, ultimately, the central concern of language understanding. Similarly, modular theories of planning may well determine which goals and plans to pursue based on an assessment of *other* goals that the agent has, but they do not generally take into account other, externally determined, features of the situation in which the agent finds itself. Thus, for example, in deciding what plan to use, such a planner will take into account potential conflicts with other goals. What it tends not take into account, however, are the resources and opportunities that are actually available in the current circumstances.

Fundamentally, then, these modular theories are based on a "noninteractive" model of knowledge: Information can interact freely with other information of the same type -- i.e., concerned with the same subject matter and couched in the same representational vocabulary -- but its interaction with other kinds of information is practically nonexistent. Mental processing is viewed as a collection of relatively isolated specialists defined, not in terms of their function, but in terms of the kinds of knowledge they manipulate. In contrast, integrated processing is based on the hypothesis that knowledge of different kinds, concerned with different content, derived from different sources, and couched in different vocabularies, can and should interact in the service of making rational decisions. To the extent that mental processing is viewed as a collection of specialists, those specialists are defined, not in terms of the kinds of knowledge they manipulate, but in terms of their functional role in processing.

In sum then, an integrated approach to planning and understanding is one which aims to produce models that attempt to avoid or minimize backtracking by making decisions about where to allocate their effort on a rational basis. In order to be as rational as possible, furthermore, such decisions must, to the greatest extent possible, employ any information that might be useful, from a wide variety of sources if necessary.

In understanding, this relevant information must include the goals and hypotheses of the understander. Thus, in an integrated approach the interpretation of an input is determined, in part, by such contextual elements. However, this alone does not suffice to make a model of understanding integrated. Even in a modular theory of understanding, the goals and hypotheses of the understander may ultimately play a role in determining whether or not a

given interpretation is appropriate. They will not, however, affect the *manner* in which that interpretation is produced. Contextual influences are viewed as a filter, applied to a potential interpretation only after it has been constructed. In an integrated approach, the goal of minimizing backtracking argues that such contextual influences should be felt at the earliest decision points that they possibly can. Thus, the understander's goals and hypotheses must play a role in determining not only the ultimate interpretation of an input, but in determining *how* that interpretation is arrived at. In particular, in an integrated theory of understanding, the inferences that are drawn in attempting to understand an input are determined in part by the context in terms of which that input will be understood. In comparison to a modular theory, an integrated theory of understanding will be a relatively goal-directed, or *top-down* , theory.

Similarly, in an integrated theory of planning, the information relevant to making decisions must include situational features external to the planner. Such contextual elements must therefore play a role in determining which goals and plans will be pursued. But again, this alone does not suffice to make a model of planning integrated. It is also necessary that such contextual influences be felt as early in the planning process as possible. That is, the external situation in which a planner finds itself should play a role in determining not only which goals and plans will be pursued, but in determining *how* those goals and plans are arrived at. In other words, in comparison to a modular theory, an integrated theory of planning will be a relatively data-driven, or *bottom-up,* theory.

2. Search in planning

In the overwhelming majority of planning situations, there are many alternative plans that are potentially useful in accomplishing any given goal. This is beneficial in that, to the extent that a planner is aware of these different plans, it is more flexible and thus has a better chance of constructing a viable plan. The flexibility offered by knowing alternative plans, however, is a mixed blessing: It immediately gives rise to the problem of somehow choosing among the alternatives offered. For example, if planning is conceived as a hierarchical process -- in which goals give rise to plans, which in turn give rise to other goals (subgoals of the original goal), and so on -- then the combination of all possible choices gives rise to a very large space of possible plans, only a small number of which may actually be feasible in the current circumstances. Thus, even leaving aside the formidable problem of interactions between the sub-goals, a hierarchical planner must come to grips with the problem of searching through a combinatorially explosive space of possible plans.

Not surprisingly, in many planners for limited domains, this problem has simply been

finessed in one of two ways. Often, the planner is given very little choice, or even no choice at all, in determining which plan to try for a given goal, because only one or two plans are known to it. This strategy, however justified within a given domain, gives rise to the illusion that the problem of search has somehow been solved, since the plans devised by such a planner are in fact devised without appreciable search (see, e.g., Sacerdoti, 1977, which along with its impressive achievements suffers rather severely from this defect). But this is in reality just a manifestation of what McDermott (1981) has dubbed the "wishful control structure fallacy." Because the planner is not given any alternative plans for most goals, it lacks the flexibility that such options offer, and hence will only work on a small number of problems. That is, the search space has been limited by fiat of the programmer to include only those plans needed to solve the particular problems at hand. What one has in these cases is not a general theory of planning, but rather a description of the correct choices necessary to pursue a small set of goals and plans in limited circumstances.

What makes the use of this strategy seem almost ironic in these cases is that it undercuts the functional considerations that motivate a theory of general purpose planning in the first place. If a given application domain requires only a limited number of plans, then there is no need to invoke anything like a general purpose planner to construct those plans. Indeed, there may not be any need to *construct* the plans at all -- they might well just be pre-computed and indexed in a table under the appropriate goals. What makes a theory of general purpose planning attractive is the possibility it offers of constructing a planner that is flexible -- that can use its knowledge to build new plans. If search has been finessed, that can only be at the expense of this kind of flexibility.

The only alternative, then, is to bite the bullet and perform some kind of search through the space of possible plans in order to construct one that does the trick. However, even the adoption of this approach does not entail confronting the problem, given a limited enough domain. If the search space is small enough -- and, in many theoretical examples, even if it is not -- it is a common strategy to simply rely on back-up to enumerate the set of possible plans until one is found which does the trick. That is, given a set of options, such planners simply make an arbitrary choice, and are prepared to back up and take another alternative if their guess doesn't work out (see, e.g., Tate, 1977).

There are two serious objections to such an approach, however. First of all -- and this point has been well understood since the earliest days of AI -- the fact of the matter is that the use of blind search is not practical as the number of choices is expanded, and if the set of viable solutions is sparse enough, because of the size of the search space that results. And, once the

protection of working only within a limited domain is removed, most planning problems turn out to exhibit these characteristics.

Second, to use blind back-up is to admit that, although one's theory might specify *what* the possible answers to a problem are, it does not specify any particular *process* for solving the problem. That is, it does not specify *how* to find a correct answer, other than by enumerating them and checking whether or not each works. Given that one of the chief aims of artificial intelligence is to develop such process models, such an admission is, to say the least, disheartening.

What is needed in a theory of general purpose planning, then, is the ability to make choices between alternative plans in a non-arbitrary fashion, rather than arbitrarily. But what could count as a grounds for making such a decision? Obviously, such top-down criteria as *a priori* likelihood of success, difficulty, etc., will be relevant. Much more important, however, are contextual features which make one or the other plan seem more likely to succeed *without* undue effort. That is, one should choose a plan in part because the situation facilitates its pursuit.

Part of this situational context includes, of course, the other goals that are being pursued, and interactions among them. It can be quite difficult simply to produce a plan which is internally coherent in the sense that the actions which serve one subgoal do not interfere with another subgoal, and much work on planning has been concerned with how to take such "internal" contextual factors into account. But just as important, and perhaps even more important, are simply the external, and quite possibly uncontrollable, factors of the situation in which the planner finds itself, and current models of planning take rarely such factors into account. At best, they are capable of noticing that some precondition for a plan is already satisfied in the situation, and are thus able to avoid the unnecessary effort of planning to achieve that precondition. But they cannot, for example, choose a plan on the grounds that one of its preconditions has already been satisfied.

In an integrated approach, such relevant information about the external situation in which the agent currently finds itself (or can expect to find itself when the plan is executed) must be taken into account in choosing appropriate plans for pre-existing goals. Plans should be chosen or rejected in part on the basis of what opportunities the situation offers, for example, based on what resources are currently available. Moreover, the earlier that such contextual information is taken into account in the planning process, the more subsequent choices its application will eliminate. By narrowing the available choices in this way, the search for an

internally consistent plan is simplified as well, because the space in which that search is conducted is made smaller. In other words, the task of finding a consistent plan which is also feasible will be made easier by taking feasibility into account as early as possible.

In sum then, we can see that for a planner to avoid the use of back-up, the planner must make use of whatever knowledge is available, in particular, knowledge of the external situation in which it finds itself, when deciding what lines of reasoning to pursue. The external situation in which the planner finds itself should, in a sense, exert a bottom-up influence on the process of planning. That is, reducing back-up in planning requires integrating top-down and bottom-up processing. Integration has an additional dividend as well: It enables a planner to be *opportunistic* -- that is, to recognize and seize unforeseen opportunities in the world.

3. Language understanding, descriptive theories, and search

The problems posed by the need to choose among alternatives is, if anything, more acute in language understanding than in planning. Because of such problematic characteristics of language as lexical and structural ambiguity, vagueness, ellipsis, and metaphor, an astonishing number of choices are available as to the proper interpretation of a given fragment of input. Further, which choice is taken will affect similar choices later in the input, so that the combinatorial problems which arise in hierarchical planning arise in understanding as well. And, just as in planning, one of two alternative strategies is typically employed to deal with this problem.

The first strategy is to artificially limit the choices available to a language analyzer to exactly and only those necessary in order to analyze the target inputs. As before, this can be justified exactly to the extent that the goal of the project is simply to accomplish the task within a limited domain. For example, a data-base application may limit context sufficiently that only one sense of most words need be considered. However, such an approach succeeds in eliminating search only at the cost of sacrificing flexibility: An analyzer and lexicon designed for one data-base will not work for another. This is the case, for example, with the customized lexicon employed by the analyzer (Birnbaum and Selfridge, 1981) for Kolodner's (1984) CYRUS model of conceptual memory.

The alternative -- which has been far more widely accepted in language understanding than in planning, for some reason -- is to make such choices arbitrarily, while remaining prepared to back-up and select other alternatives should they prove mistaken. This is the strategy that has been adopted by most syntactic analyzers, for example augmented transition

network parsers and descendent models based on Prolog. In the last section, however, I argued that such an approach simply evades the issue of how a mental ability should be implemented, and substitutes instead a naive enumeration of all possible solutions, in this case, all possible syntactic analyses of inputs. Even within the syntactic analysis community, there has been a growing realization that stronger claims must be made about the processes involved in language understanding (see, e.g., Marcus, 1980).

In fact, when applied to language analysis, the indiscriminate use of arbitrary decisions and backtracking is, in many ways, even less satisfactory methodologically than when applied to planning. The reason is as follows: In order to develop a serious process theory for some task, that is, a theory in which choices are not made arbitrarily, one needs a theory of *what* information will be relevant in making choices, and *when* it will be relevant. But in order to construct such a theory, one first needs to establish *what* the choices are, and those choices must be such that sources of information can in fact be found that will help in making them rationally. That is, the theory of what choices need to be made must be such that a theory of *intelligent* choice among the alternatives can be developed. An *arbitrary* theory of the choices which must be made will not necessarily suffice.

In planning, the choices which must be made are, at least, justifiable on intuitive grounds. That is, we can examine our intuitions as to whether or not it seems plausible to consider some goal or plan in the service of some other goal or plan. In language analysis, however, no such intuitive justification can be produced for the set of choices to be made, since we are not consciously aware of, for example, alternative ways of referring, or of constructing a noun phrase -- assuming we do such things in the first place.

Put more generally, the problem we come to is this: A theory which is capable of taxonomically describing the range of appropriate behavior in some domain in terms of a series of choices can be trivially turned into a "process model" that generates such behavior by the use of arbitrary choice and the ability to back up. Whether it can be turned into a *serious* process model -- one that makes empirical claims beyond those offered by any veridical description of the behavior -- depends on whether ways can be found to make the choices non-arbitrarily, and thus dispense with the indiscriminate use of backtracking. But we can expect that there are, in general, many alternative ways to characterize the set of decisions involved in some mental ability. There is no reason to believe, given any one such characterization, that some way can be found to make the choices so characterized in a non-arbitrary way. That is, there is no reason to believe that such a *descriptive* theory need be embodied in the final *process* theory, or even that it would be particularly useful in constructing a process theory, especially if the

behavior that it characterizes -- e.g., grammaticality judgments -- is functionally rather peripheral.

This is by no means intended as an attack on the utility of descriptive theories, but rather on their misapplication. There are many important descriptive results in cognitive science. Psychology has developed sophisticated experimental techniques to uncover hidden aspects of human performance. Linguistics and anthropology have developed sophisticated methods for probing people's intuitions, and sophisticated notations for the results. But none of these fields is concerned with bringing functional considerations to bear in constructing process models of mental abilities: That job falls to AI.

Of course, a great deal of good work in AI has a certain descriptive flavor to it as well. For example, representation theories concerned with developing good vocabularies for expressing the content of our knowledge in a given domain -- to take two rather disparate examples, Schank and Abelson's (1977) theory of the representation of plans, goals, and themes in mundane situations, and Hayes's (1985) theory of the representation of the naive physics of liquids -- are often descriptive. (For eloquent defenses of the importance of such content theories, see Hayes, 1979, and Newell, 1982.) The point here is that a straightforward implementation of these theories in a way that relied on arbitrary choice and backtracking, for example in a theorem proving system, would add nothing to them.

Thus, simply translating Schank and Abelson's theory into Horn clause form and implementing it in Prolog would add nothing to their work, at least from a scientific perspective. The result might be a program that understood simple stories, but it would not be a process model of story understanding. Similarly, an implementation of Hayes's theory of naive physics in Prolog would be just that: an implementation. The result might be a program that solved problems in naive physics, but it would not be a contribution to the theory of problem-solving. Or, to cite a recent contribution to the literature, straightforwardly implementing a previously developed linguistic theory of conjunctions in Prolog, as Fong and Berwick (1985) do, adds nothing whatsoever to that linguistic theory. It may be useful, but it makes no additional empirical claims beyond those made by the linguistic theory alone. Such a program cannot, therefore, be considered a *bona fide* process model of how coordinate expressions are understood. (For a biologist's view of the need to distinguish true process theories from descriptive theories, see Sydney Brenner's discussion of how this issue arises in understanding the genetic control of morphological development, quoted in Judson, 1979, pp. 217-221.)

This entire argument is, in fact, a generalization of one of the standard objections to theories of syntactic analysis. What basis do we have for believing that people actually use syntactic rules of the sort devised in linguistics -- descriptive rules -- in constructing syntactic representations devised for a similar descriptive purpose -- or that any language understander should? Do we even make the choices that such theories imply? Can a way be found to incorporate such choices into a serious process model of language analysis -- one that does not rely on back-up? If a way cannot be found to construct a modular theory of syntactic analysis without arbitrary choice and backtracking, then it is a methodological imperative to look for a kind of analysis which *can* be accomplished without them.

It is a methodological imperative for the following reason: If arbitrary choice and the concomitant indiscriminate use of back-up are allowed, then if it is possible to construct a computational model at all, it will always be possible to construct one that is modular in the sense discussed above. That is, the claim that some mental ability can be accomplished in a modular fashion is *irrefutable* if arbitrary choice and back-up are permitted. Whenever a decision seems to require contextual information of the sort that an integrated theory would attempt to take into account, a modular theory could be constructed that simply chose arbitrarily, and was prepared to back up if that choice proved mistaken.

Exactly the same argument holds for planning. Thus, if a theory of planning in some domain seems to require arbitrary choice and back-up, then the set of choices being presented to the planner is probably wrong and we should look for another. For example, one might construe the planning problem in chess as being the choice of one move over others. However, this approach clearly entails a great deal of search and back-up. Therefore, it is probably better to look for another way to characterize the planning problem in chess, in which the choices to be made are not merely between one move and another.

In sum, it seems equally crucial to eschew the use of arbitrary choice and backtracking in both planning and understanding. Planners must try to arrive at appropriate goals and plans on the basis of rational decisions. To be rational, then, those decisions must take into account the factors that affect the appropriateness of goals and plans. In part, the appropriateness of goals and plans depends on the extent to which they can be successfully pursued, and successful pursuit in turn depends on external conditions, such as the availability of resources. So, to control search, an integrated planner must use its knowledge of the conditions that hold in a situation, current or expected, in deciding how to arrive at goals and plans that will be appropriate.

In language understanding, the story is much the same. Understanders must try to arrive at the appropriate interpretations of input utterances, again on the basis of rational decisions. To be rational, those decisions must therefore take into account the factors that affect the appropriateness of such interpretations. That depends, in part, on the extent to which the interpretation is coherent in the context, which in turn depends on whether it is explicable in terms of the understander's goals and hypotheses. So, to control search, an integrated understander must use this contextual information in determining how to arrive at an appropriate interpretation.

4. Outline of thesis

In this chapter, we have stated our problem: Arbitrary, unmotivated decisions in a process model of some mental ability are a sign that we do not really understand how that ability is accomplished, or why it should be accomplished that way. Except when blind search is inescapable -- for example, in cases where the agent is genuinely ignorant -- its use cannot be permitted in a serious process model. Nevertheless, most AI theories of planning and understanding depend, either explicitly, or what is worse, implicitly, on arbitrary choice and the ability to back up.

Some readers may wonder whether the situation is as grim as I have portrayed it here. The answer to that question can be seen in the AI applications programs -- primarily expert systems and natural language data-base query systems -- that are currently undergoing commercial development. The narrow range of competence of such programs, and the rapidity with which their performance degrades as they are pushed to their limits, is a widely acknowledged fact. These undesirable characteristics are largely due to the fact that the designers of these systems have had little choice but to sacrifice flexibility in order to avoid combinatorially intractable search, for our knowledge of how to bring relevant information to bear on the decisions that such systems must make -- or even of what that information *is* -- remains rather primitive. For example, to take an application with which I am somewhat familiar, natural language data-base query systems typically give short shrift to the problem of lexical ambiguity. Because such systems are devoted to a particular application domain, many meanings of many words can simply be ignored, since they are usually -- although not always -- irrelevant to the domain of interest. The ambiguities that cannot be ignored are resolved on the basis of *ad hoc* tests that are tuned to the specific domain, and which therefore will not work properly in another context. That such systems can nevertheless be made useful is a tribute to clever problem choice and backbreaking programming efforts.

In order to avoid either the inflexibility of current applications systems, or the equally undesirable alternative of blind search, programs must make decisions on a rational basis, using whatever knowledge and information seems relevant. Thus, the process of determining which goals and plans to pursue must take into account, among other things, features of the environment in which the planner finds itself. The process of arriving at an interpretation for some input in a given situation must take into account, among other things, the understander's current hypotheses about that situation.

However, the ability to assemble and apply knowledge of disparate types and from disparate sources in the service of rational decisions about which lines of reasoning to pursue cannot itself be expected to be cost-free. Thus, integrated processing will be justified only to the extent that it reduces the amount of arbitrary search required in planning or understanding by more than its own cost. There is reason to be concerned that the cost of gathering and weighing the information relevant to making a decision rationally might outweigh the cost of making an incorrect choice. The fear that this is in fact the case is what gives modular theories much of their attraction.

This is particularly true with regard to modular theories of understanding. Although it is well established and uncontroversial that understanding requires plausible inference of a surprisingly sophisticated sort, the apparent complexity of this inferential processing seems to have convinced many that its use should be avoided at all costs (see, e.g., Woods, 1973). From this perspective, an integrated approach which seeks to apply such inferential processing as early as possible in understanding appears counterproductive. The sort of argument that seems to motivate this view goes something like this: "Understanding is fast and relatively error-free. It almost has the feel of an act of perception. Inference -- well what that really means is problem solving, and problem solving is a slow and clunky business, not to mention unreliable. So understanding can't involve much inference, except maybe at the very end."

There are, it seems to me, several mistakes in this line of reasoning. The most basic, although least consequential, is the misapplication of phenomenological evidence. It is true that understanding has the quick, sure feel of an act of perception. But if anything is clear by now from thirty years of AI research, it is that such intuitions are in no way an indication of underlying simplicity. Perception itself is hardly a simple act. There is, furthermore, no reason to believe that inference, or even problem solving, must be slow and error-prone. It is true that most current *models* of inference and problem solving are slow and error-prone. And it may be that problem-solving, particularly in the absence of good knowledge, really is sometimes as awkward as advertised. But inference, and even problem solving, can often be

just as fast, and feel just as much like an act of perception, as understanding does. In a domain with which we are familiar, we often seem to "see" the answer to a problem almost as soon as it is posed.

More substantively, however, I believe that fears about the cost of integrated processing are based on a serious misconception as to the average number of choices (i.e., the branching factor) that actually obtains at real decision points, and about the number of such decision points, in real problems in planning and understanding. As Minsky (1963) points out, under such circumstances it is worth devoting a substantial effort to reasoning which can direct inference along productive paths. Although it cannot be said that cost is no object in the effort to make decisions rationally, the cost of not trying is almost surely too high.

The key issue in achieving integrated processing is how to make the relevant information available to the decisions that require it. In the simplest cases, exactly what information is relevant may be known ahead of time. In such cases, the link between that relevant information and the rule that makes the decision can be, in a sense, "hard wired." That is, the decision rule can be formulated to take certain factors -- and perhaps *only* those factors -- into account in making its decision. For example, the use of a particular class of plans for some goal may require certain critical resources, the availability of which is not assured, and for which no alternatives may be substituted. In that case, it seems clearly useful to make the decision as to whether to attempt to construct a plan based on some member of that class contingent on the availability of the crucial resources. Indeed, if the necessary conditions for the pursuit of some goal or plan are only intermittently available, it may pay to consider whether to pursue that goal or plan *whenever* the relevant conditions happen to arise -- that is, it may pay to be an *opportunistic* planner.

In general, however, there is no way of knowing, ahead of time, exactly what information will be relevant to the decision -- in a new context, a new factor may be significant. Thus, the methods by which relevant information is made available to a decision must be flexible enough so that new factors can be taken into account in new contexts. For example, the decision as to the appropriate interpretation of an ambiguous word cannot be made -- as I will show in chapter four -- merely on the basis of some fixed, foreseeable set of conditions. In such cases, an integrated approach must attempt to specify methods by which pathways can be dynamically created between decisions and the information that is relevant to making them rationally. As we will also see in chapter four, this has all too often not been the case.

Indeed, the shortcomings in prior work on integrated processing described in chapter

four is one of the motivations for the work described here: Integrated processing is a functionally well-motivated idea, but constructing process models which truly adhere to spirit of those functional motivations has proven difficult. In part, I believe this is because the functional considerations that motivate integrated processing, and the requirements that those functional considerations impose on process models, have not been specified with sufficient clarity. That is one of the chief aims of this thesis.

Thus, in part I of this thesis, which concerns understanding, I will address the issue of how to bring contextual information, in particular information arising from the understander's hypotheses about the situation it is trying to understand, to bear on decisions that arise early in the interpretation of an utterance in that situation. The particular type of decision with which I will often be concerned is the interpretation of vague or ambiguous lexical items. Chapter two will discuss the relationship between syntax and semantics from an integrated point of view. Particular attention will be focussed on the problem of how to dynamically integrate constraints arising from different sources in order to facilitate rational decisions in language analysis, particularly lexical disambiguation. Chapter three presents a critique of the foundations of syntactic analysis -- with which most modular theories of language understanding are, almost obsessively, concerned -- from a functional, AI perspective. Chapter four presents an analysis of the problem of lexical ambiguity, and of the shortcomings of previous approaches, which motivates an integrated approach to the problem. It also demonstrates how and why previous attempts to apply such an integrated approach have generally failed to fulfill their promise. Chapter five moves to the larger issue of controlling inference in language understanding, with particular application to the problem of applying abstract thematic structures in story understanding.

Part II of the thesis concerns integrated processing in planning. I will show that the kind of attention to conditions in the external world which seems necessary to make rational decisions and avoid back-up leads to a conception of planning much closer in spirit to that which seems appropriate in everyday human life. In particular, I will outline the basic ideas of *opportunistic* planning, discuss their applications, present the issues which must be addressed by a theory of opportunistic planning, and sketch out a rough taxonomy of the ways in which those issues might be resolved. Ultimately, this investigation will lead us to a surprising convergence with Freudian psychology.

CHAPTER 2

INTEGRATING SYNTAX AND SEMANTICS

1. Introduction

The relationship between content and structure in natural language is unquestionably one of the most confusing issues in cognitive science, and apparently one of the most durable as well, since it arises repeatedly in the literature on psychology, linguistics, and philosophy of language. From the perspective of artificial intelligence, this relationship can best be understood by considering the functional roles that each plays in achieving the goals of language use, primarily communication. In particular, this means that the relationship between content and structure is to be understood not only in terms of the information that each contributes to the processes of understanding and generating language, but also in terms of how and when that information is applied to resolve the problems that naturally arise in the course of attempting to carry out those processes.

In an integrated approach to language understanding, as we saw in the last chapter, as much information as possible -- including, in particular, the understander's goals and hypotheses -- should be taken into account in making rational decisions as early as possible in the understanding process. In other words, considerations of *content* -- meaning, world knowledge, and context -- are presumed to play a role quite early in language analysis (see, e.g., Schank, Tesler, and Weber, 1970; Schank, 1975; Riesbeck, 1975; Riesbeck and Schank, 1978; Schank, Lebowitz, and Birnbaum, 1980; Schank and Birnbaum, 1984). This position, which I will call the *integrated processing hypothesis*, stands in direct contrast with modular theories of language understanding, which assume the existence of a logically autonomous syntactic analysis procedure, preceding, and providing input for, semantic and inferential memory processing (see, e.g., Fodor, Bever, and Garrett, 1974; Marcus, 1980). This chapter is concerned with exploring the implications of the integrated processing hypothesis, the problems that it must address, and its relation to other approaches.

Two observations led to the hypothesis originally (Schank, 1975). The first was the failure of syntax-oriented approaches to the construction of systems for processing natural language, particularly in the early research on machine translation (see, e.g., Bar-Hillel, 1960).

This failure was due primarily to problems of ambiguity and implicit content, and considerations of meaning and context are crucial to the solution of both sets of problems. Thus, a more content-oriented approach to language analysis seemed necessary.

The second observation was the rather commonsense one that it is easier to understand a foreign language, especially when reading, than to speak or write it. Most of us have had the experience of picking up a magazine written in a language with which we are slightly acquainted, and more or less understanding it, especially if we know something about the topic. Paraphrasing or answering questions in the language, however, would be beyond our capabilities. It would seem, then, that understanding language is not nearly as dependent on syntactic information as generation is. Indeed, even in our native language we often seem to vary how we approach the task of understanding, depending upon the material to be understood, our interest in and familiarity with that material, our goals, and other contextual factors. The use of syntactic information does not seem an all-or-none proposition.

The above considerations led to the development of a series of *expectation-driven* language analyzers employing the sort of basic semantic knowledge captured by conceptual dependency theory (Schank, Tesler, and Weber, 1970; Riesbeck, 1975; Riesbeck and Schank, 1978; Gershman, 1979; Birnbaum and Selfridge, 1981). In these analyzers, expectations arising from unfilled case roles in incomplete conceptual structures are used to guide the assembly of a complete representation of an input utterance as a whole. They have proven moderately successful in a variety of settings, including story understanding, question answering, and dialog systems.

Simultaneously, the development of memory structures for representing the pragmatic world knowledge necessary in order to understand, such as scripts (Schank and Abelson, 1977), led naturally to the question of how the information embodied in such structures might be used in language analysis. DeJong's (1979) design for an integrated understanding system -- the first attempt to utilize such memory structures in parsing -- resulted in a program, FRUMP, that applied simplified scripts directly to the problem of skimming and summarizing newspaper stories. Although skimming a story is not nearly as difficult as arriving at a deep understanding of it, FRUMP's success at this task provided some tangible evidence that the way to solve the problems of language analysis is to bring as much knowledge as possible to bear as early as possible in the understanding process.

Perhaps the best motivation for an integrated approach to language analysis, however, stems from an appraisal of its role within a functional conception the language understanding

process as a whole. In a nutshell, understanding crucially involves relating what you hear or read to what you already know, assimilating it in such a way that any new information can be recalled and employed when relevant. Thus, memory is intimately related to the very properties of communication that make it so useful and important. Understanding an input involves, in part, finding the most relevant memory structures available to explain it, and creating for the input a new memory structure derived from the old ones (see Schank, 1982). In particular, then, language comprehension depends crucially on knowledge about the situations which language describes. The common sense motivation for integrated processing is that it doesn't make sense *not* to make the fullest possible use of this information when it is relevant.

One of the most important functional applications for such information is in resolving ambiguity. That people do in fact use world knowledge in disambiguation during the course of sentence analysis can be demonstrated by considering garden path sentences, in which the use of such knowledge leads them astray. For example, in the course of processing the sentence "The old man's glasses were filled with sherry," (Schank, 1973), most people incorrectly decide that "glasses" means "eyeglasses," and are, as a result, surprised by the rest of the sentence. They are consciously aware of having made an error and of repairing it. This error can only be explained by assuming that in understanding this sentence, people apply their knowledge of the relationship between age and eyesight to resolving the ambiguity of "glasses" *before* completing any putative low-level analysis of the sentence, since otherwise the context provided by such an analysis would presumably be available to provide the correct interpretation. Indeed, if "bartender" is substituted for "old man" in this sentence, the opposite interpretation is made in processing, and so no back-up is needed.

2. What is the problem?

The issue of how content and structure are related in language processing can be divided into three closely related, but nevertheless distinct, questions. These questions are often conflated, and the failure to keep the distinctions clear has been a major factor inhibiting mutual comprehension among those who have argued the issue.

The three questions correspond to three aspects of any computational process that can be usefully distinguished. The first aspect of interest is the *control structure* of the process, which defines the sub-processes that must be invoked in order to accomplish the task, the information that must be supplied to them as input, and can be expected from them as output, how they communicate, other information to which they have access, and the conditions under which

they should be invoked. The second aspect is the *representational structures* that are constructed and operated on by the process, and that constitute the inputs and outputs of any sub-processes and of the process as a whole. The final aspect is the *knowledge base* of information about the domain which the process uses in the performance of the task. Since this knowledge often takes the form of rules, the knowledge base is commonly called a *rule base*. Our three questions, then, concern what an integrated approach to language analysis, as opposed to a modular approach, might possibly entail with respect to these three aspects of a language processing system.

> The first question concerns the processes which apply the information used in language understanding: Are information about content -- meaning, world knowledge, and context -- on the one hand, and structural information, on the other, applied by separate control mechanisms, or are they applied jointly by the sub-processes that constitute understanding? That is, does language understanding proceed by the independent application of each type of information in succession, by modules devoted solely and exclusively to one type of information, or by their joint application in modules which have access to a wide variety of information in carrying out their functions?
>
> The second question involves representational structures: Are the structures that are used to encode information about content separate from those used to encode information about structure? That is, does language understanding involve the computation of an independent level of syntactic representation? What kinds of representations must be constructed in order to understand?
>
> The last question concerns the knowledge used in understanding. Can the rule base that embodies syntactic knowledge be isolated from that which embodies semantic and pragmatic knowledge? Is there a clean separation between these sets of rules, or is there a continuum of rules, some concerned solely with content, some concerned solely with structure, some concerned with both?

Although it is important to distinguish these questions, the possible answers to them are of course interdependent, as will be seen below.

2.1. Question one: Control structures

> The question of whether information about content and information about structure are

applied jointly by the sub-processes that constitute understanding, or by independent sub-processes, is of course the question of whether or not an autonomous syntactic analysis procedure exists. The claim that both should be applied jointly follows rather directly from the basic tenets of an integrated approach to understanding, namely, that contextual influences should play a role in rational decision-making as early in the understanding process as possible. In a sense, then, this is the weakest possible version of the integrated processing hypothesis. In particular, if it were not true that syntax and semantics were applied jointly, it would be difficult to argue for either integrated representational structures or an integrated knowledge base in language processing. The opposite position -- the claim that syntax and semantics are applied by independent control mechanisms -- is the strongest possible form of syntactic autonomy.

It is instructive to map out some of the possible alternatives concerning the issue of integrated versus independent control processes for applying information about content and information about structure to the problem of language analysis. What follows are sketchy descriptions of several possible positions.

[a] *Syntax and semantics are completely separable.* By this position, syntactic analysis is a completely independent process, logically and temporally prior to the content-based inference processes involved in understanding. This position implies that syntax alone controls language analysis at the earlier points in processing. This is the view that results when the descriptive model outlined in Chomsky (1965) and descendent models are given a straightforward, if naive, recasting into the performance domain.

[b] *Syntax and semantics are "nearly decomposable."* By this view, it is still the case that a syntactic analysis process precedes semantic processing, and provides the input for it. However, this process may on occasion query a semantic component in order to make a syntactic decision. This limited interaction between content-based inference processes and the syntactic analysis process is controlled by syntax, in that only the syntactic mechanism can decide that some interaction is required. This is the position taken by Fodor, Bever, and Garrett (1974) with their theory of independent syntactic processing within clauses, and by others -- e.g. Woods (1970), Kaplan (1975), and Marcus (1980), among others -- with somewhat more flexible communication regimes between syntactic and semantic processing.

[c] *Syntax and semantics have a "heterarchical" relationship.* By this position, information about content, on the one hand, and structural information, on the other, are still applied by separate control processes. However, their relationship is construed as far more

cooperative than in the preceding position. In a sense, the syntactic processor and content-based processes operate somewhat like co-routines. That is, the interaction is no longer exclusively under the control of the syntactic mechanism. A syntactic component does some work, then calls some content-based process which does what it can and then in turn calls syntax for more information, and so on. This appears to be the position advocated by Winograd (1972 and 1977) and Lytinen (1984). This description might in fact apply to models that also fit the preceding or following descriptions, since "independence" is a fuzzy concept: As the richness and frequency of communication between modules increase, the modules become more integrated and less independent.

[d] *Syntax and semantics are applied jointly, in an integrated fashion.* By this view, information about content and information about structure are both employed, jointly, in the process of language analysis, and whatever available information seems most useful will be applied in the rational resolution of the problems that arise in determining the appropriate interpretation of an input. This is the position taken here, and previously advocated by Schank, Tesler, and Weber (1970), Riesbeck (1975), Riesbeck and Schank (1978), Schank, Lebowitz, and Birnbaum (1980), and Riesbeck and Martin (1986). A similar view seems to inform the experimental program being carried out by Marslen-Wilson and his colleagues (see, e.g., Marslen-Wilson, Tyler, and Seidenberg, 1978). One way in which this highly integrated class of models might be differentiated from the heterarchical class described above is that whereas the latter models construct and operate on explicit syntactic representations (as in, e.g., Winograd, 1972), highly integrated models -- for instance, those cited here -- may well not. We will return to this point in the next section.

One point of clarification seems necessary here. The opposition on the above spectrum between position [a], logically independent syntactic analysis, and position [d], joint application of syntactic and semantic information in integrated control structures, is often misconstrued. In particular, since position [a] implies that syntax alone controls early language analysis, the opposite position is often taken to be something like "semantics alone controls early language analysis." This is clearly wrong, but that in no way affects the validity of position [d], which doesn't imply anything of the sort.

2.2. Question two: Representational structures

In the course of understanding or generation, any language processing system must compute some structures for representing content, on the one hand, and words and their properties, on the other. An important question, then, is what additional structures must be

computed to represent the structural information associated with utterances? There are basically two positions that one can take.

[a] *An autonomous level of syntactic representation, such as phrase markers, must be computed.* For example, Fodor, Bever, and Garrett (1974, p. 368) claim that "the structural analyses to be recovered are ... precisely the trees that a grammar generates," by which they mean that in the course of comprehending language, a language understanding system, and in particular, people, must compute explicit syntactic representations of the sort employed by generative linguistics in describing syntactic phenomena.

[b] *No independent level of syntactic representation is constructed or operated on during language processing.* This claim has an important consequence for models of language processing: Whatever structural distinctions need to be represented, must be represented either at the level of conceptual structures, or at the level of words. An important implication of this claim is that if a conceptual representation carries structural information about an utterance that is necessary for subsequent processing, then that information must also serve some semantic or pragmatic function. In other words, any elements added to a conceptual representation for the purpose of carrying structural information must be justified independently in terms of some conceptual function. (Interestingly, this last point bears a mirror-image resemblance to an observation made by Katz and Postal, 1964, in arguing that syntactic transformations must preserve meaning. As they pointed out, this claim implies that any difference between the meanings of two sentences must be reflected in some difference between the syntactic deep structures underlying those sentences. Further, they recognized that in order to support the original hypothesis, they as theorists were required to justify such differences in deep structure on independent syntactic grounds.)

This question of representational structures has an important bearing on our earlier discussion of control structures. If two processes acted on, and produced as output, different sorts of structures, such behavior would constitute one characteristic that would lead us to say that the two processes were independent of each other. Hence, without the computation of explicit and independent syntactic representations in language understanding, one of the characteristics that might lead us to single out an independent syntactic processor would be missing. In this sense, the computation of an independent level of syntactic representation is a weak prerequisite for the existence of an independent syntactic processor. Hence, claims of autonomous syntactic processing are always accompanied by the presumption that independent syntactic representations are necessary. While an integrated approach to language analysis does not necessarily entail denying this presumption, the argument for integrated processing

would be strengthened if it turned out to be false. I will, therefore, provisionally take the stronger position here, and assert that information about content and information about structure are represented, jointly, by integrated representations.

2.3. Question three: Rule base

The strongest possible form of the integrated processing hypothesis would be the claim that there are no purely structural rules -- that is, that all the rules used in language processing refer, in some way, to information about content. If this were true, then the claims of integrated representations and processes would follow immediately. The opposite claim -- that purely syntactic rules indeed exist -- is the weakest possible form of the hypothesis of syntactic autonomy. Without such a set of purely syntactic rules, for example, the claim that there is an independent syntactic processor doesn't even make sense. That is, the extent to which syntactic knowledge can be separated from knowledge of content determines whether an independent syntactic analysis procedure is even logically possible.

There is little support for the strongest possible form of the integrated processing hypothesis. Indeed, as far as I know, no one has ever claimed that purely syntactic rules do not exist. The existence of purely syntactic rules does not, however, entail the need for independent syntactic representations or an autonomous syntactic analysis procedure to apply those rules. While freely acknowledging the existence of purely syntactic rules, and their utility in language analysis, the integrated processing hypothesis does make the claim that these rules simply occupy one extreme of a continuum of rules, and are not distinguished *by use* from other sorts of rules. This follows from the two prior claims of the integrated processing hypothesis that (1) language processing is effected by the joint application of information about structure and content by integrated control mechanisms, and (2) no independent level of syntactic representation is computed in language processing. If these two claims are true, then purely syntactic rules are not functionally distinguishable by use from other sorts of rules. Thus, the integrated processing hypothesis is supported to the extent that the role of purely syntactic rules in processing can be shown to be similar to the role of rules concerned with other sorts of information.

2.4. The integrated processing hypothesis

We are now in a position to state exactly what the integrated processing hypothesis claims:

First, it claims that language understanding proceeds by the joint application of information from different sources, and concerned with different content, in the rational resolution of the problems that naturally arise, rather than as a collection of processes characterized by their sole concern with one particular source and type of information. This is in contrast to the models proposed by Fodor, Bever, and Garrett, Woods, and Marcus, among many others.

Second, it claims that no independent level of syntactic representation is constructed, operated on, or output by the language analysis process. This is in contrast to all of the above models, as well as the model proposed by Winograd.

Third, it claims that purely syntactic rules -- purely syntactic in the sense that they are expressed in a vocabulary concerned only with structural concepts -- are not used differently from other sorts of rules. That is, they are functionally integrated in processing and play no privileged role. This follows from the first two claims.

3. Some psychological evidence

There has been a great deal of psychological experimentation bearing on the relationship between syntax and semantics in language analysis. In this section I will review a few results that seem to support the integrated processing hypothesis. Among psychologists, the results of this work have convinced even the strongest partisans of generative linguistics that:

1. There is no evidence that people make use, in comprehension or generation, of the kinds of rules devised by generative linguists to describe syntactic phenomena.

2. The very strong claim of a completely autonomous syntactic processor (position [a] in section 2.1 above) cannot be upheld.

In fact, these points constitute the most "conservative" interpretation of the experimental results, in the sense of conserving some role for generative linguistics in psychology. Less sympathetic observers will note that the results, although consistent with various patched-up claims of syntactic autonomy, were not as predicted by theorists who advocate that position.

One of the earliest and most significant results was uncovered by Slobin (1966). Using a picture verification task, he investigated differences in how long it takes to understand passive sentences as compared to active forms, distinguishing between "reversible" and

"irreversible" passives. A reversible passive is a sentence like "John was seen by Bill," in which syntax must be consulted to determine who saw whom. That is, the meaning of this sentence can only be distinguished from that of "John saw Bill" by attending to the fact that one utilizes a passive construction while the other does not, since both are equally sensible. An irreversible passive is a sentence like "The ice cream cone was eaten by John," in which, by virtue of semantics, one can determine who ate what. That is, this sentence can be distinguished from "The ice cream cone ate John" on the grounds that the latter is semantically anomalous. What Slobin found is that, although reversible passives take longer to understand than the corresponding active forms, irreversible passives do not.

It is tempting to conclude from this result that the human language analysis mechanism makes no use of syntax unless semantic information alone seems insufficient. However, Slobin's results are not so unequivocal. Part of his study included presenting subjects with sentences which appear perfectly sensible without any use of syntax, but which are in fact semantically anomalous if syntax is taken into account. An example is a sentence such as "The boy was raked by the leaves." If syntax were simply ignored, then this sentence could be understood as meaning "The boy raked the leaves," which is perfectly sensible. Nevertheless, subjects usually detected the anomalous nature of these sentences. On the other hand, their error rates on this kind of material were much higher than usual -- that is, they often *didn't* notice the anomaly. The main conclusion to be drawn from Slobin's results, then, is that the functional relationship between syntax and semantics in human language processing is a complicated one. This is what one would expect if they were involved in a highly integrated system. In theories which are based on some notion of syntactic autonomy, on the contrary, the relationship between syntax and semantics is quite simple and straightforward: That is the whole *point* of such models.

Marslen-Wilson and his colleagues have performed numerous experimental studies on the empirical status of autonomous syntactic processing. A representative result can be found in Tyler and Marslen-Wilson (1977). They studied the model proposed by Fodor, Bever, and Garrett (1974), a chief claim of which is that, within clauses, sentence analysis proceeds by the operation of a completely autonomous syntactic processor, and in particular, no higher-level knowledge can enter the process until a clause boundary is reached. To study this claim, subjects were presented with sentence fragments such as the following:

1. If you walk too near the runway, landing planes...
2. If you've been trained as a pilot, landing planes...

They were then immediately supplied with a probe word, either "is" or "are," and asked to simply repeat the probe word as quickly as possible. On pragmatic gounds, as determined by the content of the first clause, "is" is appropriate as a continuation of second fragment but not the first, whereas "are" is appropriate as a continuation of the first but not the second. The only way to determine whether a probe is appropriate is on the basis of meaning and pragmatic knowledge, making use of the context created by the content of the initial clause of the test sentence fragments. The data showed that subjects were slower to repeat an inappropriate probe. Since the appropriateness of the probe is a syntactic property -- number agreement -- and since subjects were probed in the middle of an uncompleted clause, this result demonstrates that whatever syntactic processing is going on is not independent of semantic and inferential processing, even within clauses.

Shwartz (1980) examined several possible structural strategies -- some proposed by him, some proposed elsewhere in the literature -- for determining pronominal referents, some of which depended on the existence of explicit syntactic representations, and some of which did not. The study found no evidence for the use of strategies which depend on explicit syntactic representations. This result strengthens the integrated processing hypothesis because it demonstrates that an aspect of understanding that might have been thought to depend on explicit syntactic representations in fact appears not to.

One final study I will mention concerns an investigation into the putative independence of semantic processing and pragmatic inferential processing in language understanding. Gibbs (1979) investigated a claim by Clark and Lucy (1975), among others, that understanding indirect speech acts requires computing, in a fairly bottom-up fashion, the "literal meaning" of the utterance, which then serves as input to pragmatic interpretation rules in order to uncover the speaker's intended meaning. Clark and Lucy had shown that, in the absence of any context, indirect speech acts did take longer to comprehend than, for example, direct requests. This was taken as evidence that an extra processing step was being performed, presumably involving the application of the pragmatic rules to the previously computed "literal meaning" of the input. Gibbs performed a similar study, in which, however, the indirect speech acts were embedded in a suitable context. He found that, in context, indirect speech acts take no longer to interpret than direct language; he thus called into question the claim that the "literal meaning" of an utterance as a whole must be computed in language understanding.

4. Functional integration of syntactic rules

The functional utility of structural information stems from the fact that the use of

semantic information alone is often incapable of producing the complete interpretation of an utterance, or, what is worse, leads to an erroneous interpretation. For example, in an utterance like "Mary gave John the magazine," it is structural information that enables the understander to determine who gave the magazine, and who received it. Syntactic knowledge, then, is primarily concerned with how to determine the roles of constituent concepts in the representation of an utterance on the basis of their positions (or rather, the positions of the words to which they correspond) in that utterance -- or, in many languages, on the basis of explicit markings attached to those words. Such information is necessary whenever the semantic representation of an utterance contains several roles that have overlapping semantic requirements. For example, both the **Actor** and the **To** case roles of an **ATRANS** (representing the concept of transfer of an abstract property) can appropriately by filled by a "higher animate." Syntactic information must, therefore, be used to decide which of several appropriate roles such an entity should be assigned.

An important feature of integrated language analysis is the *functional* integration of syntax and semantics. In part, this means that rules concerned solely with content, rules concerned solely with structure, and mixed rules should all play similar roles in processing. We can illustrate this point with a few examples. The functional conception of syntax sketched in the preceding paragraph is that syntactic knowledge is necessary when semantics alone is insufficient to correctly determine the meaning of the input, such as when the semantic restrictions on roles in a conceptual structure are not unique -- that is, are not mutually exclusive. This suggests that when a semantic role does have unique semantic requirements, no syntactic knowledge may be necessary to find the correct entity for that role. Hence, that entity might have an extremely free syntax with respect to the entire construction.

It turns out that one can find examples of this sort of phenomenon. Consider the conceptual object of **MTRANS** (representing the concept of communication). An act of communication takes an entire conceptualization, or proposition, as its **Object**, namely the concept being communicated. Since, on semantic grounds, no other case role of an **MTRANS** can be assigned a complete proposition, it is possible that the position of the **Object** concept might vary rather freely with respect to the overall **MTRANS** construction -- and, while the **Object** proposition itself imposes certain syntactic restrictions, on the whole it does show considerable variability in position. Consider the following examples:

A Liberian tanker ran aground off Nantucket Island, the Coast Guard said.

The Coast Guard said a Liberian tanker ran aground off Nantucket Island.

A Liberian tanker, the Coast Guard said, ran aground off Nantucket Island.

A Liberian tanker ran aground, the Coast Guard said, off Nantucket Island.

A Liberian tanker ran aground off Nantucket Island, said the Coast Guard.

These examples suggest that the rule used to determine the **Object** of an **MTRANS** conceptualization can, at least sometimes, be based entirely on semantics. It simply looks for an entire conceptualization to fill the **Object** case role.

On the other hand, relative subclauses -- which might be initiated by a relativizer such as the word "that," as in, for example, "The car that I saw in the showroom..." -- require a great deal of syntactic information to be properly analyzed. Roughly speaking, the analysis of relative subclauses entails the use of a rule like the following:

> To the right will be found a conceptual structure with some unfilled role(s). Use the concept to the left to fill (one of) the role(s), in accordance with semantic and syntactic requirements. Then take the resulting conceptualization and subordinate it to the concept on the left.

This rule is purely syntactic: It refers to relative positional information and to "*unfilled roles*," but it says nothing about the kind of concept or about restrictions on roles. This is not to say that the rule can necessarily be used without reference to content, however, since such information may need to be consulted to produce a meaningful subordinate conceptualization. The important point here is that this purely syntactic rule plays exactly the same role in processing as does the purely semantic rule for determining the **Object** of an **MTRANS**. Syntactic and semantic knowledge cannot be distinguished by use in an integrated model of language analysis.

This same point can be made by considering generation, for instance of noun groups. A purely syntactic rule that one might want to have in a generator is that adjectives precede nouns. But the order of the adjectives themselves is not determined by such a purely syntactic rule. A generator must have enough knowledge to know that "big red ball" is generally more appropriate than "red big ball," or that "old Irish grandmother" is more appropriate than "Irish old grandmother." Examples of this sort can best be explained by postulating that an adjective supplying information about a more "intrinsic" property should be closer to the noun than one supplying information about a less "intrinsic" property (Clark and Clark, 1977). The proper generation of noun groups depends crucially on the simultaneous application of both this rule and the purely syntactic rule that adjectives precede nouns. But the notion of "intrinsic

property" is clearly conceptual, not syntactic. Thus, the problem of generating noun groups properly is another example that argues for the functional integration of purely syntactic rules with other sorts of rules.

5. Using inferential memory in language analysis

The motivation for an integrated approach to language analysis stems from the possibility it offers of utilizing context and world knowledge to rationally guide early decisions about the proper interpretation of an input, and thus avoid the need to make such decisions arbitrarily. The decisions that arise early in the interpretation of linguistic input include, for example, the proper resolution of lexical and structural ambiguity, of vagueness, of anaphoric reference, and so on. Thus, an integrated approach to language analysis can only be justified to the extent that it is able to bring context and world knowledge -- in particular, the understander's hypotheses about the situation it is attempting to understand -- to bear on the rational resolution of such problems in language analysis. It is, therefore, crucial for integrated models to address this issue: The failure to do so would simply undercut the reasons for adopting an integrated approach in the first place.

The first problem that arises in attempting address this issue is one of communication: How can contextual information and world knowledge be made available to decisions for which they are relevant? It is this problem that provides the functional motivation for integrated representations. To the extent that the representations operated on by different sorts of rules are of the same type, interaction between them is facilitated. Indeed, such communication will be the easiest when the representational structures employed by inferential memory and by language analysis are not just of the same type, but are actually the same *structures*. Thus, one of the key approaches to integrated language understanding is to analyze input utterances directly into the memory representations that organize world knowledge and facilitate inferential understanding (Schank, Lebowitz, and Birnbaum, 1980).

In some cases -- when the meaning of some word in the utterance directly asserts the relevance of the appropriate memory structure -- this is relatively unproblematic. For example, consider the sentence "Bruno kidnapped Lindbergh's baby boy." The word "kidnap" points directly to a memory structure (let's call it **M-KIDNAP**) which contains the following scenes: The actor takes control of the victim; hides him; contacts the victim's relatives and demands a ransom; negotiates a deal with the relatives; picks up the ransom; releases or kills the victim; and tries to elude capture. The best representation of this sentence must include an instantiation of this memory structure, in which the roles of **Actor**, **Victim**, and **Relatives** are filled by

Bruno, the baby, and Lindbergh respectively. Indeed, if the understander possesses prior knowledge of the Lindbergh baby kidnapping, then the best representation would be in terms of the previously instantiated memory structures representing such knowledge. In either case, the instantiated memory structures must additionally represent the fact that the first scene, in which the actor takes control of the victim, has been accomplished. This would enable the understander to recognize the anomaly of a sentence like "Bruno kidnapped Lindbergh's baby boy but failed to grab him."

The important point for an integrated approach to language analysis, however, is that by immediately attempting to represent the utterance directly in terms of the **M-KIDNAP** memory structure, all of the knowledge about kidnapping which that structure organizes can be made available to solve problems in language analysis. Thus, for example, the fact that other scenes of the structure are available to provide expectations explains why, in a sentence like "Bruno kidnapped Lindbergh's baby boy and left a note," we so easily understand that the note is a written document -- as opposed to a musical note -- that it is probably for Lindbergh, and that in fact it instantiates the second scene of **M-KIDNAP** -- contacting the relatives and demanding a ransom.

In other words, the most important property of a memory structure such as **M-KIDNAP** as far as language analysis is concerned is the fact that the semantic requirements it imposes on other structures to which it might bear some relation -- which is to say, the *expectations* to which it gives rise -- are extremely detailed and specific. This has several immediate consequences. First, as the semantic requirements imposed by a representational structure become more specific, the chance that those requirements will overlap -- that is, that a given entity can satisfy more than one of them -- can generally be expected to decrease. It follows, then, that if some entity meets the semantic requirements, for example, of some case role in a high-level memory structure, and if those requirements are specific enough, then that is quite likely the appropriate role to assign it. Thus, to the extent that the more specific semantic constraints arising from specific memory structures can be brought to bear in language analysis, the functional rationale for utilizing syntactic information is reduced. It is for this reason that the IPP program (Schank, Lebowitz, and Birnbaum, 1980; Lebowitz, 1980) is able to analyze fairly complex sentences using far less knowledge of syntax than might be expected. Indeed, this is also the explanation for our ability to read and understand text about a subject with which we are familiar in a language which we do not know well enough to speak or write, or when skimming.

Perhaps more importantly, however, the more detailed and specific the semantic

constraints that are available, the more likely they are to be useful in solving problems in language analysis such as lexical ambiguity. Thus, the expectations associated with specific memory structures such as **M-KIDNAP** are likely to be particularly useful in this regard. One way -- probably the simplest -- that such information can be brought to bear is in the course of attempting to determine the assignment of case roles in such structures. Of course, this will be true on *any* account of language understanding. The challenge for an integrated approach is to bring such constraints to bear as early as possible in the understanding process. Thus, an integrated model of understanding must attempt to assign case roles in specific memory structures as directly as possible, so that the semantic constraints associated with those roles can be applied to problematic linguistic inputs as early as possible.

There are some cases in which this is relatively straightforward. For example, we know that the **Victim** of **M-KIDNAP** is most likely to be a person who is very dear to someone who has a great deal of money. Indeed, in a given instance of this structure we may know exactly who the victim is, to whom he or she is related, and how. We also know that the direct object of the verb "kidnap" should be assigned to fill the **Victim** role. Thus, we know that the semantic constraints associated with the **Victim** role should be applied to the direct object of the verb "kidnap." The correspondence is direct simply because the verb "kidnap" points directly to **M-KIDNAP**.

There are more interesting cases, however. Consider the verb "demand." Within the context of a story about a kidnapping, if the kidnapper is the one doing the demanding, then the direct object of this verb should be assigned to a very specific case role in **M-KIDNAP**, the role of **Ransom**. If that correspondence is known, then the assignment can be made *directly*. This, in turn, will ensure that any semantic constraints associated with the **Ransom** role can be brought to bear immediately on the direct object of the verb "demand." These constraints include, in particular, that the **Ransom** will be something that the kidnapper considers valuable, often just money.

Now, consider the sentence "The kidnappers demanded more dough." The word "dough" has (at least) two possible meanings, one being "a mixture of flour and water (and possibly other ingredients) used to bake bread, pastries, etc.," and the other being, colloquially if somewhat archaically, "money." If the correspondence between the **Ransom** role of **M-KIDNAP** and the direct object of the verb "demand" is known, then the constraints associated with the **Ransom** role can be applied to immediately disambiguate "dough" as meaning, in this case, "money."

Thus, an integrated approach to language analysis depends on the *joint* application of syntactic constraints -- such as "direct object of the verb" -- and specific, detailed, semantic and contextual constraints -- such as those arising from case roles in high-level memory structures. The key problem here is this: How are these constraints combined? That is, how is the correspondence between the **Ransom** role of **M-KIDNAP** and the direct object of the verb "demand" established? When we read a fragment of a sentence like "The kidnappers demanded...," we know that what comes next should be assigned to the **Ransom** role of **M-KIDNAP**. Such an expectation yokes together syntactic knowledge about an entity -- knowing that it comes next in the sentence -- with knowledge from a specific, high-level memory structure -- knowing that it fills the **Ransom** role of **M-KIDNAP**, and should therefore fulfill all of the semantic constraints pertaining to that role. How can these two sorts of knowledge be put together?

In the case of a word like "kidnap," of course, they are already together. That is, because the word "kidnap" points directly to the memory structure **M-KIDNAP**, the correspondence between the direct object of the verb, on the one hand, and the **Victim** role, on the other, can be presumed to have been established and stored in memory when the word "kidnap" was first learned. Can we assume that the same is true of the word "demand" as well?

Perhaps, but not without a bit of complication. The problem is that the word "demand" can be used in many contexts other than kidnapping stories, and in those contexts its direct object should not be assigned to fill the **Ransom** role of **M-KIDNAP**. Indeed, the semantic constraints associated with that role would probably be entirely inappropriate in such cases. For example, in the sentence "This recipe for strudel looks better, but it demands more dough," it should be clear that the intended meaning of "dough" is not "money." Thus, if we assume that the correspondence between the direct object of "demand" and the **Ransom** role of **M-KIDNAP** has already been established and stored in memory, we must in effect assume that the word "demand" is *ambiguous*, with one meaning specifically applicable in the context of kidnapping stories, and another meaning (or several other meanings) applicable in other contexts.

In the case of a word like "demand," this is not entirely implausible. It seems reasonable to assume that words which are used repeatedly in a particular way within a particular context will eventually develop a specialized meaning within that context. For example, it also seems likely that the word "order" has a specialized meaning within the context of restaurant stories, a meaning which is not unrelated to the meaning it has more generally -- "command" -- but

which is nevertheless distinct.

However, this solution is inadequate for the general case. Instead of the verb "demand," for example, a story about a kidnapping might use the phrase "say they want," or innumerable other locutions. In fact, *any* locution describing the expression of some desire on the part of the kidnappers in a kidnapping story is probably intended to refer to their demands -- that is, to the **Ransom** role of **M-KIDNAP**. It is extremely unlikely that the correspondence between the **Ransom** role of **M-KIDNAP** and the syntactic role of its linguistic realization in *every possible* locution that can be used to express desire in English -- or, for that matter, in other languages -- has been determined previously and stored in memory.

Nevertheless, if an integrated approach to language processing is to succeed in any but the simplest cases, some way must be found to establish the correspondence, so that the specific semantic constraints associated with case roles in specific memory structures can be brought to bear, as early as possible, on the relevant linguistic input. For example, in the sentence "The kidnappers said they wanted more dough," the disambiguation of the word "dough" should be accomplished just as immediately and directly as if the verb "demanded" were used.

In sum, even a relatively simple aspect of integrated language analysis -- the early and direct application of specific semantic and contextual constraints derived from case roles in specific memory structures to problematic linguistic input -- requires combining information of diverse kinds and from diverse sources. In the very simplest cases, such information may already be combined in a single expectation, so that the problem reduces to one of retrieving that specific expectation under the appropriate circumstances. In general, however, such a *static* approach to the integration of different kinds of knowledge in language analysis is inadequate. What seems necessary is a more *dynamic* approach.

In fact, there is an even more basic problem which must be confronted here. The difficulty with a word such as "demand" or a phrase such as "say they want" is that, unlike a word such as "kidnap," they do not necessarily directly assert the relevance of some specific memory structure. That is, the word "kidnap" can be expected to point directly to the **M-KIDNAP** memory structure. But many words and phrases -- perhaps most -- cannot be expected to point directly to such structures. In particular, the phrase "say they want" and even, in all probability, the word "demand," does not point directly to the relevant sub-parts of the **M-KIDNAP** memory structure. Rather, the determination of how they relate to the context of **M-KIDNAP** seems to result from the attempt to explain their role in that context.

In other words, the proper interpretation of a "demanding" action in the context of a kidnapping, or any other context, cannot be determined simply by *retrieving* the appropriate sense of "demand:" It must be *inferred*.

6. Vagueness, ambiguity, and flexible integration

Let's consider another example in which the proper interpretation of a sentence in terms of the appropriate high-level memory structure must confront this sort of difficulty:

Joe bought his new TV at Macy's.
Joe got his new TV at Macy's.

The first utterance, by use of the word "bought," directly asserts an instance of the "buying" memory structure (henceforth, **M-BUY**). Hence, in an integrated approach it is best represented by instantiating **M-BUY**, with Joe as the **Buyer**, Macy's as the **Seller**, and the TV as the **Goods**. Since the word "buy" points directly to **M-BUY**, this is not at all difficult to achieve. The proper representation of the second utterance is not quite so straightforward, however. Because memory representations should as much as possible reflect similarities in meaning (Schank, 1975), then, since these two utterances are synonymous, or nearly so, their representations should be identical, or nearly so. Therefore, the second utterance should also be represented by an instance of **M-BUY**.

That could be accomplished easily if the word "get" pointed directly to **M-BUY** in the same way that "buy" does. However, since there are many ways to get something besides buying it, this approach would imply that "get" is an extremely ambiguous word, with innumerable subtly different senses. If "get" were unique in this regard, this consequence would be tolerable; but it is not. Perusing any text yields words with the same characteristic, words like "take", "use," "go," "have," "cut," "send," "carry," and so on. In fact, this technical problem of an explosively large number of distinct word senses -- which Rieger and Small (1979), among others, have argued must be addressed, and have then attempted to solve -- arises because the entire approach remains, at root, based on the old notion that the meaning of an utterance is a simple, additive function of the meanings of the words it contains. Using such an approach, if one took **(M-BUY Buyer (Joe) Seller (Macy's) Goods (TV))** as the representation of "Joe got his new TV at Macy's," and then subtracted out the disambiguated meanings of all the words in that utterance, nothing would be left over. Every nuance of the utterance, every subtle distinction in the meaning of every word in the context of the utterance, must be reflected in one of the innumerable number of precomputed senses of the

words themselves. Again, such an approach seems too static. The appropriate interpretation of "get" in this context should be determined not by merely retrieving a previously computed sense of the word, but by some sort of inferential processing.

The alternative I propose is based on the intuition that the problem with a word such as "get" is not that it is enormously ambiguous, with many possible meanings of great specificity but rather that it's meaning is *vague* and general. What "get" conveys is simply a *crude description* of what "get" might mean in a given context. In order to derive a more highly elaborated and specific representation for the utterance as a whole (in this case, in terms of **M-BUY**), some kind of inferential processing must be employed in the attempt to explain what a "getting" action entails in the particular context. (The importance of crude descriptions as a starting point for understanding was discussed in general terms by Marr, 1977). In the simplest cases, such an explanation could be provided in a manner akin to script application, by matching a portion of some contextually active memory structures (Schank and Abelson, 1977; Cullingford, 1978). In a sense, such an explanation process works simply by using the crude description as a search key for indexing inside of such contextually active memory structures.

To see how this might work, let us assume that the meaning of "get" is represented simply as **ATRANS** (representing the concept of transfer of possession or control). Further, assuming that one knows that Macy's is a department store, **M-BUY** would be potentially relevant, since department stores are a common setting for **M-BUY**. **M-BUY** contains several scenes, among them two which center around instances of **ATRANS**: One represents the transfer of the goods from seller to buyer, and the other the transfer of money from buyer to seller. Now consider the following informal explanation rule:

> If an action occurs in a setting which is commonly associated with some activity, then look inside the memory structure that organizes knowledge about that activity, and check whether the action could instantiate one of its scenes. If so, then instantiate the entire memory structure, and mark the matching scene as already accomplished.

Since the transfer of the TV to Joe matches a central scene in **M-BUY** -- the transfer of *the* ***Goods*** *from the* ***Seller*** *to the* ***Buyer*** -- by using a rule of this sort, the utterance "Joe got his new TV at Macy's" can be understood as an instance of **M-BUY**, without necessitating that "get" point to **M-BUY** as a possible sense. In particular, this approach does not limit the meaning of an utterance to be simply an additive function of the meanings of the words that make it up. If one subtracted away the meanings of the words in this example, one would be

left with an instance of **M-BUY** which, although *suggested* by "Macy's," was not directly *asserted* by it or any other word in the utterance.

Utterances that require the sort of processing described above are extremely common. I will present one more example here:

John mailed me a postcard from Mexico.
John sent me a postcard from Mexico.

As above, these two utterances are synonymous, or nearly so. Both are best represented in terms of the memory structure that represents our knowledge of postal service (i.e., **M-MAIL**). In the first case, this is directly asserted by the use of the word "mailed." In the second case, it must be inferred, since the word "sent" has been used instead.

"Send," like "get," is a vague word: It points to a crude description of what it might mean in some context. Let's assume that this meaning can be represented simply in terms of **PTRANS** (representing the concept of physical transfer of location). It seems clear that the word "postcard" suggests that **M-MAIL** might be relevant. Since the main goal of **M-MAIL** is to accomplish the **PTRANS** of some object, and since the action asserted by "send" is a **PTRANS**, it is fairly straightforward to conclude that **M-MAIL** should be instantiated.

The sort of processing described above can also solve our original problem of deriving the proper interpretation of utterances involving the word "demand" or the phrase "say they want" within the context of a kidnapping story. Suppose that the meanings of these locutions are represented simply in terms of the **MTRANS** (representing the concept of communication) of some desired goal. (This is not quite right, since the use of the word "demand" additionally implies the potential to carry out some threat, but it is good enough for our present purposes.) Suppose further that the **M-KIDNAP** memory structure includes a scene in which the kidnappers communicate their demands to the victim's relatives, represented in terms of a similar **MTRANS** structure. We have, therefore, something like the following structures in memory:

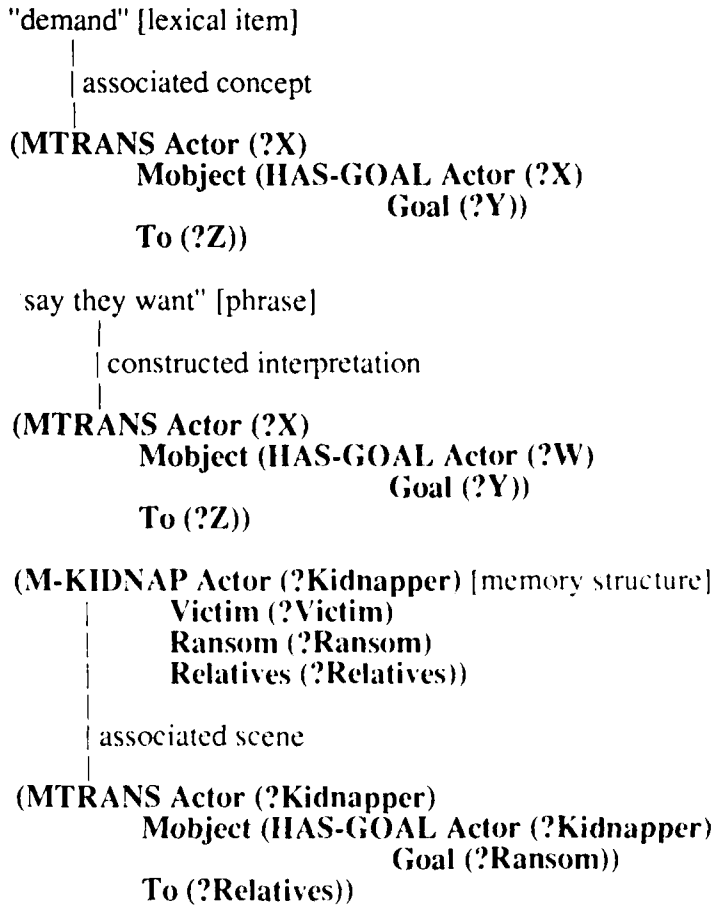


Figure 2-1: Memory structures for interpreting demanding actions in kidnapping context

Now, in the course of attempting to explain an input such as "The kidnappers demanded..." or "The kidnappers said they wanted..." in the context of a kidnapping story, the information associated with **M-KIDNAP** -- that is, the expectations that it generates -- can be used, as described above, in a manner akin to script application. Thus, for example, both of these inputs could be explained because they match the expectation for a scene in which the kidnappers communicate their demands to the relatives of the victim.

The crucial point here is that this use of memory structures in the explanation of linguistic inputs can form the basis for a more flexible solution to the problem of integrating information of different kinds and from different sources. The reason is as follows: In the course of the kind of pattern-matching that such an explanation process entails, the filler of the **Mobject** role of the **MTRANS** arising from "demand" or "say they want" must be unified with the filler of the **Mobject** role of the **MTRANS** in the matching scene of **M-KIDNAP**.

(As Charniak, in press, has pointed out, the form of unification involved here is not the standard one, because variable role fillers such as **?Ransom** cannot be construed as universally quantified variables.) In particular, the two **Goal**s must be the same. Thus, variable role fillers **?Y** from the input and **?Ransom** from the matching scene in the memory structure must be presumed to be the same if the input is to be explained in terms of this scene. But, of course, **?Ransom** is just another name for the filler of the **Ransom** role in **M-KIDNAP**. Thus, the filler of the **Goal** role in the input **MTRANS**, **?Y**, must be the filler of the **Ransom** role in **M-KIDNAP**.

At this point, all of the constraints bearing on both the **Ransom** role of **M-KIDNAP** and the **Goal** role of the input **MTRANS** can be seen to be jointly bearing on exactly the same entity. The constraints bearing on the **Ransom** role are, of course, the specific expectations that **M-KIDNAP** represents about what kidnappers want. What are the constraints bearing on the **Goal** role of the input? They are, simply, the syntactic constraints that specify where in the utterance its filler can be expected. For example, if the input were conveyed in terms of the word "demand," we would expect the direct object of the verb to be the **Goal**. In short, in the course of attempting to explain the input in terms of the **M-KIDNAP** memory structure, the correspondence between the direct object of the verb "demand," or, for that matter, the phrase "say they want," on the one hand, and the **Ransom** role of **M-KIDNAP**, on the other, can be *dynamically* established.

Once that correspondence has been established, the joint application of contextual and syntactic constraints required for integrated processing is made possible. However, that alone is not sufficient: An integrated approach to language analysis must also attempt to bring such constraints to bear as *early* as possible. Thus, the sort of explanation process described above must take place in "real time," during the course of sentence analysis. In particular, in the example discussed above, the correspondence between the direct object of the verb and the **Ransom** role of **M-KIDNAP** must be established *before* the direct object is actually read or heard, so that the relevant contextual constraints can be applied immediately in the resolution of any linguistic problems it may pose. In this way, the flexible integration of different sorts of knowledge, from different sources, that is required for an integrated approach to language analysis, can be achieved.

7. Conclusion

One reason to expect that artificial intelligence can contribute to our understanding of the relationship between syntax and semantics is that it must address the issue in order to construct

process models capable of performing the types of linguistic tasks that people can perform. Other approaches to linguistic theory are under no such methodological pressure to address the issue; indeed, quite the opposite is true. Research aimed towards elucidating a competence theory of syntax, for example, naturally starts by de-emphasizing the relationships between content and structure. This essentially *methodological* decision has often been transmuted into *empirical* claims for various forms of syntactic autonomy, usually without any consideration being given to what sort of relationship might be required in order to actually perform significant language processing tasks. In research on artificial intelligence, however, this lack of attention to functional requirements is -- or ought to be -- impossible.

The motivation for an integrated approach to language analysis lies in the potential utility of the contextual information embodied in explanatory memory structures to early decisions in the analysis process, in particular the resolution of such problems as lexical ambiguity. However, there has been a great deal of confusion about what such an approach actually entails. In particular, it does *not* mean that language analysis does not involve the use of syntactic information, or that information about structure cannot be distinguished from information about content. The issue is not, and never has been -- despite claims to the contrary (see, e.g., Marshall, 1980) -- whether sentences are different from other objects in the world, or whether facts about sentences -- e.g., facts like "word x precedes word y in sentence z" -- are different from facts about other things -- e.g., facts like "If patron x places order y with waiter z, then z will bring y to x." These claims are trivially, indeed tautologically, true and without empirical import. For AI, the issue that is in contention is whether or not the useful and efficient *analysis* of linguistic inputs can best be accomplished by utilizing information about their content, and about the larger context in terms of which they must be understood. For linguistics, the issue is whether or not facts about sentences can ultimately be explained -- or even *described* adequately -- without reference to such information.

In order to apply contextual information stemming from the understander's hypotheses about a situation to the rational resolution of problems in early language analysis -- thereby avoiding, as I argued in chapter one was necessary, the need to make such decisions arbitrarily -- "communication channels" must be formed to make the relevant contextual information available to the processes that make those decisions. That is, the representations which constitute the understander's hypotheses, and those which arise from the text, must be related to each other in such a way that constraints from both sources can be merged and taken into account simultaneously. The timely disambiguation of a word, for example, depends on knowing that it might fulfill some role on account of its position in the sentence, and also that that role must meet certain specific semantic requirements in the current context. In the

simplest cases, as we saw, the necessary connections between an hypothesis and a text might already exist in memory, easily retrievable because the text contains words or phrases that refer directly to the hypothesis. In general, however, this is not the case, and inferential processing must be employed in order to determine how the input is related to the hypothesis, thereby dynamically establishing the connections necessary for integrated processing.

CHAPTER 3

THE FOUNDATIONS OF SYNTACTIC ANALYSIS: A FUNCTIONAL CRITIQUE

1. Introduction

A non-integrated view of a mental faculty such as language understanding must start by breaking the task into a series of relatively non-interacting modules, each defined by its concern with a single type of knowledge from a single source. Although there are many non-integrated theories of semantic and inferential processing -- as we will see in the next two chapters -- by far the largest single category of theories based on a non-integrated approach to natural language understanding are those concerned with the syntactic analysis of sentences. I have already criticized such theories obliquely in the last two chapters. In this chapter, I will present a more direct critique of modular theories of syntactic analysis.

The primary aim of this critique will be to show where and how modular theories of syntactic analysis are deficient *as theories*. In particular, then, their merits or deficiencies as engineering will not be discussed. Thus, we will not be concerned here with such questions as how many constructions a given parser is currently able to analyze, how much CPU time it takes to do so, how many -- or how few, depending on how progress is measured -- rules or lexical items it employs, or the ease with which new rules can be added. Rather, the goal here is examine how well these theories address the issues which must be addressed by a computational theory of any cognitive ability. I will show that modular theories of syntactic analysis *systematically* fail to address certain issues which are of central concern to artificial intelligence, and I will argue that this failure raises questions about their standing as theories with empirical content for AI.

2. Functionality and artificial intelligence

The characteristic that most distinguishes artificial intelligence from the other cognitive sciences is its concern with *functionality*, the analysis and justification of representations and processes with respect to the functions they serve in some natural cognitive task. AI research

is based on the proposition that functional requirements, stemming from the need to perform realistic cognitive tasks, are the most important determinants of cognitive theories. This is not to say that other requirements arising, for example, from experimental results, or from accurate descriptions of human behavior, should not play a role in theory development, even in AI. In linguistics and psychology, of course, these are the primary sorts of requirements that theories must attempt to satisfy. But these motivations alone are not sufficient, and are not even necessarily central, to artificial intelligence. They do not distinguish AI from either linguistics or psychology: Function does.

An AI theory of language understanding, then, must attempt to define and justify the representations and processes it proposes with respect to their functional role in the understanding process as a whole. Nevertheless, modular theories of syntactic analysis, almost without exception, fail to provide functional justifications for even the output that they claim must be produced, let alone the processes by which they propose to produce that output. This systematic failure reveals that these theories are not, at least as currently formulated, functional theories at all. Instead, they are more or less straightforward implementations of the descriptive theories constructed by linguists concerned with syntactic competence. As I pointed out in chapter one, however important such descriptive theories may themselves be for cognitive science, their implementation adds nothing to their empirical content, although it may prove useful from an engineering point of view. It is thus not surprising that many linguists view AI work on natural language processing as a relatively uninteresting engineering adjunct to their science (see, e.g., Dresher and Hornstein, 1976).

Let's examine this point more closely, to see exactly why the straightforward implementation of a descriptive theory is so unproductive. To provide some neutrality to the discussion, I will construct a hypothetical case. Consider a behavior like *reminding* (Schank, 1982), in which an input situation or story elicits some memory. Based on a simple and universally acknowledged premise -- that memories are retrieved because of features that they share with inputs -- Schank constructs a theory of the kinds of features that seem to serve as indices in human memory processing. Among other things, he demonstrates that the features implicated in reminding are often of a surprisingly abstract nature.

Now consider constructing a process model motivated solely by the requirement that it exhibit the sort of reminding behavior that people do. That could be accomplished, for example, by representing any given input in terms of the features that it has in common with any given memory that it elicits, and then using those features as indices to retrieve the desired memory by means of any of the commonly available indexing techniques. Such a model "gets

reminded," but then, rather oddly, does nothing with the reminding. Since the construction of such a model does not take into account *why* it would be useful to get reminded, it does not say anything about *when* it should be reminded, or *what* it should be reminded of, beyond what the original description of the behavior does. And as a result of the failure to address these functional issues, it is highly unlikely to be a correct model of *how* to be reminded -- or at least, there is no reason to believe that the methods it employs bear any resemblance to what would actually be needed in a functional memory, let alone the methods that are employed in human memory processing. Yet it is exactly these sorts of questions that an AI theory of reminding would seek to answer.

We can see the further consequences of the failure to address these issues by considering what sort of representation for memories -- which is to say, what sort of output -- such a "process model" of reminding would argue for. In fact, it would not argue for any representation at all. From the point of view of such a model, the way that episodes are represented in memory is entirely arbitrary, perhaps consisting only of the features used to retrieve them. Since the model does not address the question of why it should be reminded, there is no functional pressure on how memories should be represented.

The proper representation of memories depends, of course, on what they are used for. Thus, one of the most important questions to ask about a behavior such as reminding is what use it might serve. One possible answer in this case might be that being reminded of a previous experience while attempting to understand an input actually helps in understanding that input because the memory generates useful expectations (Schank, 1982). Another answer might be that the memory suggests potentially appropriate explanatory structures -- that is, that reminding plays a role in indexing to explanations. Yet another possible answer proposed by Schank (1982) is that you are reminded in order to be able to learn. That is, when you are reminded, you compare the current situation with the reminding in order to see if there are any interesting similarities that need to be explained. If this answer is correct -- and these answers are by no means mutually exclusive -- then the representation of the episodes in memory must be such that there is extra information in the episodes that is not contained in the structure that organizes them, since otherwise there would be nothing to learn.

The important point here is that all of the above answers imply that there is no way of knowing, simply on the basis of a description of reminding behavior, which features held in common between an input and the memory it elicits were actually used in the process of being reminded, and which were, in a sense, the "output" of that process. That can only be determined on the basis of functional utility. Thus, in any model that fails to take functional

considerations into account, the decision as to which features are used as indices, and which constitute the output of the reminding process, must be made arbitrarily. The need to make such arbitrary choices in the design of an AI theory points up the need to devote more effort to considering the goals which must be satisfied by the process with which the theory is concerned, just as does the need to make arbitrary choices within the theory itself -- indeed, the two are involved in a trade-off. Arbitrary decisions are a good clue that a descriptive theory is being implemented.

Interestingly enough, the problem of arbitrary choice in cognitive theories has not escaped notice in linguistics. Chomsky (1965), in particular, has made it a central focus of linguistic theory to find criteria which would rank one grammar as higher than another -- given that both adequately describe the same data -- in order to reduce or eliminate the arbitrariness of grammatical descriptions. However, because the doctrine of autonomous syntax precludes explaining grammatical phenomena in terms of either content or functional utility in language use, his approach to this problem is to assume the existence of a meta-grammatical level, universal grammar, which is *itself* arbitrary. In fact, the strongest psychological claim made by generative grammar -- that humans possess innate knowledge specifically concerning the structural properties of natural languages -- is based on this presumption. If the linguist's problems in determining the best description of the syntactic properties of a language are taken to be similar to those confronted by a child attempting to learn that language -- as Chomsky (1965) has argued -- then, to the extent that the linguist's problems are to be solved by appeal to the court of universal grammar, perhaps so are the child's. However, since the alleged principles of universal grammar are themselves arbitrary, there is a danger of infinite regress in this approach to the problem of language acquisition. The solution is to postulate that universal grammar is innate, language-specific knowledge. Thus, Chomsky's approach to linguistics elevates arbitrariness to the status of psychological principle.

In contrast, an artificial intelligence theory must attempt to reduce arbitrariness in the description of a cognitive system teleologically, by reference to the goals of that system. From an AI perspective, the way to reduce arbitrariness in the representation of linguistic knowledge is by considering why such knowledge is useful, and how it can best be used, in understanding and generation. Ultimately, such an approach must be grounded in the goals of language itself -- primarily, that is, in terms of how language is used to communicate meaning. There is simply no way that a functional viewpoint on language can rationally avoid this issue or its impact on the representation of linguistic knowledge. In fact, any putative rules of universal grammar uncovered by linguistics would be, for AI, not an answer to the problem of arbitrariness, but a question. Why are languages such that they have these properties, if indeed

they do? From a functional perspective, before we assume that some property of language is merely an historical accident of human evolution, we would do well to ask what use it might serve. If such a functional explanation can be found, we are no longer compelled to assume that the property is innate: The question remains open whether its utility serves as a constraint on evolution, or on learning within the individual human being.

3. Non-deterministic syntactic analysis

Probably the most widely employed method for natural language analysis is *augmented transition network* parsing, or ATNs (Thorne, Bratley, and Dewar, 1968; Bobrow and Fraser, 1969; Woods, 1970). The basic idea is to express the grammar of a language as a finite-state machine with several crucial extensions. First, the grammar is made recursive by the addition of a stack, the use of named sub-networks to represent non-terminal syntactic categories (e.g., "NP"), and the use of such non-terminal categories in addition to basic syntactic categories (e.g., "noun") to govern transitions (Thorne, Bratley, and Dewar, 1968). This is, as Winograd (1972) has pointed out, essentially adding the notion of subroutines to the grammar.

Now, this isn't quite enough to easily represent certain grammatical rules of natural languages, one of the standard examples being subject-verb number agreement in English. So the grammar must be further extended by the use of variables that can be assigned features of various constituents, which can then be compared, transferred, and so on, by tests and actions that augment the transitions (Bobrow and Fraser, 1969) -- hence the name, augmented transition network parsers. The test associated with a transition must be true in order to make that transition -- i.e., in order to traverse the arc representing that transition in the network -- and the actions are performed if the transition is made.

Such networks are used for parsing sentences by following the transition arcs, generally using depth-first search. Beginning at the "start" state, the grammar interpreter picks an arc to traverse, and attempts to find an instance of the constituent governing that arc. If the arc is governed by a terminal constituent category, the interpreter simply checks whether the next item in the input belongs to that category. If the arc is governed by a non-terminal constituent, the interpreter pushes the variables on the stack and attempts to parse an instance of that non-terminal constituent using its associated sub-grammar -- hence, the recursive nature of the mechanism. In addition, in order to successfully traverse an arc, the test associated with it must evaluate to true. If the constituent governing the arc is found, and the test is true, then the arc is traversed. The parsed constituent is removed from the input, and the action specified by the arc is performed. This usually results in building some structure representing the

constituent that has been found, and placing it in some variable.

If the type of constituent needed to traverse a chosen arc is not present in the input, or if the test on the arc is not true, the interpreter backs up and attempts to traverse another arc leading out of the current state. If none of the arcs leading out of a state can be traversed, the interpreter backs up to the previous state, replaces the constituent that was parsed in exiting that state to the input stream, and attempts to find another path forward from that previous state. In other words, ATNs are an inherently non-deterministic approach to syntactic analysis.

Thus, ATN models in no way support the claim that language understanding should or must involve the use of an independent syntactic analysis module -- at least, not if that claim is intended as an *empirical* one. For, as I pointed out in chapter one, the claim that some process can be accomplished in a modular fashion is irrefutable if arbitrary choice and the indiscriminate use of backtracking are permitted, as they are in ATNs. Indeed, the heavy dependence of ATN parsers on backtracking has been previously criticized from viewpoints as disparate as Riesbeck and Schank (1978) and Marcus (1980). Nevertheless, because the approach seems well-tailored to language analysis -- given the initial assumption that syntactic knowledge, represented in a grammar, will be the sole source of knowledge employed -- it is important to see just how general and unspecific this mechanism actually is. It is, in fact, simply a version of back-chaining -- trying to show that something is true by trying to show that its antecedents in some implication rule are true -- with back-up allowed for failed subgoals. In other words, it is simply one of the general methods by which symbolic computation can be accomplished. It is for this reason that the programming language Prolog, which implements that general method, has proven so well-suited to writing ATN parsers (see, e.g., Comerauer, 1978; Pereira and Warren, 1980).

To see that this is in fact all that is going on, consider how an ATN might attempt to parse an input sentence. First, the attempt to parse a sentence would lead to an attempt to parse a noun phrase and a verb phrase. The attempt to parse the noun phrase would lead, in turn, to an attempt to parse either a name, or else more complex phrase consisting of a determiner, followed by an adjective, and then a noun, and so on. Similarly, the attempt to parse a verb phrase would lead to an attempt to parse a verb group and then another noun phrase, and so on. This entire scenario can be characterized in terms of back-chaining as follows: To show that the input **a** is a sentence, show that it consists of a noun phrase **b** and a verb phrase **c**. To show that **b** is a noun phrase, show that it consists of a name **d**, or else a more complex phrase consisting of a determiner **e**, followed by an adjective **f**, and then a noun **g**, and so on. The attempt to parse a verb phrase could be similarly recast in terms of back-chaining.

The problem with relying entirely on such general methods, as I pointed out in chapter one, is that on this account of what it means to solve problems in AI, any problem is solved as soon as a representation theory is developed which allows us to enumerate all the potential outputs for a given class of inputs. However important representation theories are -- and they are certainly crucial -- the use of such a general mechanism to search for the correct interpretation of an input adds no empirical content of its own to those theories: It is entirely neutral.

For example, assuming that one has a representational theory in which to express the plans and goals that actors in some domain can be expected to pursue, one can implement a goal- and plan-based understanding system that explains input actions in exactly the same way, using back-chaining and back-up. To show that an actor **x** is pursuing goal **a**, such a system would attempt to back-chain and show that **x** was pursuing sub-goals **b** and **c**, and continue until it found a goal that the actor was obviously in the course of pursuing. If it turned out that one could not show that the actor was pursuing goal **a**, the system would try to see if he were pursuing goal **d**, and so on. This hardly constitutes a theory of how explanatory inferences should be drawn in language understanding. It simply implements, in the most straightforward way possible, the representational theory employed. As I argued in chapter one, a true process model must not rely on undirected search of the solution space. Rather, it must be a model of how to search the space in a directed fashion, by making rational decisions about which directions seem fruitful.

Just as one can turn an ATN into a back-chaining program, one can easily take a task which is usually viewed in terms of back-chaining -- for example, medical diagnosis -- and implement it as an ATN. One could write an ATN for diagnosis as follows. First, one would list all the diseases as arcs between the start state and the final state. Then, for each disease, there would be a sub-graph with each arc representing a symptom whose presence would be evidence for the disease. If all the necessary symptoms were present -- and, since disjunction is allowed, there could be alternative sets -- then the ATN could traverse the arc to the final state, and print out the name of the disease. If, when pursuing some arc, it got stuck, it could simply back up and check another symptom or disease. This is simply a machine for exhaustively checking all possible diseases known to the system.

To put this another way, if ATNs are a theory of language analysis, then they are also a theory of story understanding and medical diagnosis. It should be clear that they are neither of these. So, they can't be considered a theory of language analysis either. Of course, the grammar rules and structural representations which ATNs employ are themselves *bona fide*

theories, albeit descriptive. Interestingly, none of the work on ATNs actually spends much time presenting or justifying the grammar or representations employed. That's because they are more or less as proposed by linguistics. So it should be clear why this work, at least, leads many linguists to believe that AI is just engineering.

4. Syntactic representations

As I argued earlier -- and, indeed, as has been argued by AI researchers from Minsky (1963) to Marr (1977) -- representations postulated by AI theories must, above all, be able to meet appropriate functional criteria. That is, a representation theory must ultimately be justified by showing how it is useful, and a processing theory by showing how it can compute such useful representations. Indeed, the specification of these criteria forms a substantial part of any AI theory, and much research on the semantic or conceptual aspects of natural language understanding, quite varied in other respects, has been concerned with this issue.

In contrast, however, almost all theories of syntactic analysis fail to functionally define the task that they are striving to accomplish. That is, they offer no functional justifications for the sorts of structures they assert are to be produced as output -- typically, phrase markers of the sort employed by linguists describing the syntactic properties of languages -- even though the output of a computational process constitutes its most important defining characteristic.

Linguistics itself, of course, is capable of providing justifications for the use of such structures in its enterprise of describing and relating syntactic phenomena. However, such justifications, although not irrelevant, are insufficient for AI. For example, other sorts of representations that are clearly necessary for language processing, particularly strings of words and semantic representations, may be capable of efficiently performing these functions -- thus rendering explicit syntactic representations superfluous from a functional point of view -- and yet the linguist would conceivably be justified in narrowing the scope of his representations for methodological reasons. However, such methodological considerations stemming from the construction of a descriptive theory have no immediately obvious application to a functional theory. One can, obviously, proliferate representations as much as one wants from a descriptive point of view, with a different kind of representation for any arbitrary cluster of facts. From a functional point of view, however, each new type of representational structure must be functionally justified with respect to the task at hand. This means, in particular, that it must be shown that such structures facilitate accomplishing the task.

We may then wonder why it is that theories of syntactic analysis assume, without

argument, that the output that needs to be produced should consist of the sorts of representations employed by linguists. In the absence of any *theoretical* justification, we might seek a *pragmatic* one. And indeed, there is one that suggests itself rather clearly. If a syntactic analyzer employs the sorts of representations employed by linguists in their descriptive enterprise, then it will be able to employ, with a minimal amount of alteration, the rules developed by linguists in the course of describing the grammars of languages. This is, it should be clear, not a very satisfactory justification from the point of view of artificial intelligence theory.

However, an argument can be made to justify syntactic representations functionally -- even though advocates of syntactic analysis do not make it, and do not even seem to realize that one is necessary -- and it would be misleading to pretend otherwise just because syntactic parsing theorists have failed to address the issue. From a functional viewpoint, the output of a syntactic analyzer must represent information about the structural properties of an input that would be useful in understanding that input. In order to be useful in understanding, such structural information should attempt to indicate which constituents of a sentence are related, and to some extent, how -- what Charniak (1983) calls the "functional structure" of the sentence. The purpose of a syntactic analyzer, then, must be to determine the relations among the semantic constituents of an input sentence -- to determine "what goes with what" -- without actually using any knowledge of semantics.

Because use of semantic information is forbidden, the most informative representation that could be hoped for would look something like this: Take the full conceptual representation of the content of a sentence, and add a link connecting each word in the sentence to those parts of the conceptual representation to which it gave rise. Then remove all of the semantic elements of the representation -- all of the predicates, all of the role names, all of the semantic and contextual constraints, etc., leaving only the structure. What is left, essentially, is a bracketed form of the original input sentence. This is the output that should be produced by a syntactic parser.

However, syntactic analyzers actually output *labelled* bracketings of sentences. From the point of view of the internals of syntactic analysis, if structural representations are to be useful in the analysis process itself, then recurring patterns must be identified to provide expectations. Once identified, such patterns must be named in order to easily denote their relations in the rules that embody the expectations. In other words, the non-terminal symbols that label a bracketing are useful in formulating the rules used in a syntactic analyzer.

Of course, if this is to be made to work, then the expectations generated by structural representations cannot, in general, be looking for individual words, any more than the expectations in an integrated analyzer could work that way. Thus, we must assume that words can be categorized in some way, and that the expectations can be for any member of a given category. In integrated analyzers, expectations can be expressed in terms of semantic and conceptual features. In a syntactic parser, however, this would violate the original premise that content should play no role.

Thus, we must assume the existence of syntactic categories. The question then arises, what should these be? At this point, there is nothing much to say except that what everyone has decided to do is to adopt as syntactic categories the traditional parts of speech -- nouns, verbs, adjectives, prepositions, and so on. Again, it should be clear that this is a pragmatic rather than theoretically motivated solution, even within linguistics proper.

Finally, none of these arguments should be taken as implying that it is actually possible to produce even an output as devoid of semantics as that described above by means of a non-integrated, purely syntactic parser. Because natural language utterances can possess genuine structural ambiguity, it is in general impossible to correctly determine how the constituents of an utterance are related without the use of semantic and contextual information. That is, although non-determinism makes it possible to produce all legal structural analyses of an input sentence according to a given grammar, determining which one is actually correct cannot be accomplished using syntactic information alone. This issue becomes even more problematic if non-determinism is forbidden in order to sustain an empirically meaningful claim of syntactic modularity, as we will see in the next section.

5. Deterministic syntactic analysis

The most significant development in modular theories of syntactic analysis within the last ten years has been Marcus's (1980) attempt to construct a deterministic model, one that does not rely on arbitrary choice and the indiscriminate use of backtracking that inevitably results. In fact, on the view taken here, Marcus's approach to a modular theory of syntactic analysis is the only one which can possibly lay claim to any empirical significance. I want to emphasize this, because the following should not be taken as an attack on either Marcus himself or on his theory alone. Rather, my critique is focused on Marcus's work because it constitutes, in my opinion, the best to date within a modular framework.

However, I will argue below that Marcus's theory, too, systematically avoids addressing

issues that cannot properly be avoided, and that consideration of those issues militates against the success of the approach. Nevertheless, in view of his proposals, it might be argued that my earlier criticism of ATNs is somewhat beside the point. There are two reasons why this is not so. First, Marcus's criticism of non-determinism in syntactic parsing was too narrow: The fact is that the use of arbitrary choice and non-determinism leads to process models devoid of empirical significance wherever they are employed, and not only in syntactic parsing. Second, despite Marcus's critique, the increasing popularity of the programming language Prolog, however well-deserved, has resulted in the continued development of parsing models that rely heavily on back-up.

My critique of Marcus's theory will center around four sets of issues. First, I will show that the theory's failure to address the issue of genuine structural ambiguity, and more specifically the *manner* in which it fails to do so, raise serious questions about the theory as a whole. Second, I will articulate the functional requirements that must be met by the output of a syntactic analyzer in order to sustain an empirically meaningful claim of syntactic modularity, argue that the theory's failure to address those requirements begs the question, and show further that there is good evidence that in fact those requirements cannot be met. Third, I will review Crain and Steedman's (1985) empirical evidence refuting the account of garden path sentences given in Marcus's theory, and discuss the implications of these results. Finally, I will discuss the implications of the theory's failure to address the problems posed by lexical ambiguity.

5.1. Genuine structural ambiguity

At the end of the last section, I pointed out that genuine structural ambiguity poses a severe problem to a deterministic, modular theory of syntactic analysis. The dedication to modularity largely forecloses the use of semantic and contextual information to resolve such ambiguity within the language analyzer itself, while the dedication to determinism militates against generating all possible analyses automatically. What Marcus decided to do about this quandary, therefore, was to preserve modularity, but shift the burden of non-determinism to the language understanding process as a whole. In particular, he proposed that the syntactic analyzer would produce only one analysis of an input sentence at a time, even if errors were possible, and that if this analysis proved incorrect, the analyzer would be called back to the same input again, with some provision made to block the original erroneous analysis.

Such an approach does not immediately call into question the theory empirically. After all, it might prove impossible to construct a model of [illegible]

even with those ambiguities that remain. Nevertheless, it is highly questionable for several reasons. First, I think it is fair to say that such an approach greatly reduces the scope of Marcus's claims: His parser no longer needs to be concerned about resolving those potential structural ambiguities that turn out to reflect actual structural ambiguities of the input. Even more, it no longer needs to deal with any subsequent potential ambiguity that might be *caused* by prior, genuine ambiguity -- that is, potential ambiguity that might have arisen if some prior genuine ambiguity had not been resolved, expeditiously, by fiat.

Second, on this account, the rest of the understanding system must in some way be able to prevent the syntactic analyzer from producing the analysis that it originally produced -- which proved erroneous -- so that some other analysis will be produced instead. But in order to do that, the language understanding process as a whole needs some knowledge of, and access to, the internals of the syntactic module. It is, of course, exactly to avoid this kind of interaction that modular theories are proposed in the first place.

Now, what Marcus (1980) seems to imply, presumably in an attempt to forestall such criticism, is that the rest of the understanding system need not know very *much* in order to control the syntactic analyzer's behavior in this way. All that seems necessary is a communication protocol by which the rest of the understanding system could send a message indicating that the syntactic module should "reparse the input, taking a different analysis path, if the other consistent analyses are desired," (Marcus, 1980, p. 13 note 10). However, the apparent simplicity of such an interaction belies the underlying additional complexity of the syntactic module which it presupposes. In particular, in order to make such a scheme work, the syntactic module would need to keep a record of the decisions which, if made differently, would result in an alternate analysis, including its state at each of those decision points, the order in which those decisions arose, and some way of keeping track of which alternatives it had already chosen in producing previous analyses. In a word, that is, it would need all of the machinery that is needed to implement non-determinism.

The main defect of this approach, however, does not lie in the additional costs, in terms of increased complexity, that are imposed on the syntactic module itself. The real problem here is that the theory is robbing Peter, repeatedly, to pay Paul once: Determinism and modularity are preserved in the syntactic analyzer at the expense of non determinism in the language understanding process as a whole. On this account, determining the correct interpretation of genuinely ambiguous sentences requires the use of arbitrary choice and backtracking not by the syntactic module alone, but by the syntactic and semantic modules in conjunction. That is, if the knowledge of and access to the internals of the syntactic module on the part of the

understanding system as a whole are to be kept to a minimum -- as they must be, in order to uphold the original claim of modularity -- then not only can semantic and contextual processes play no role in determining the syntactic analysis of an input utterance, but the only information that they can transmit to the syntactic module about an inappropriate analysis is that it is inappropriate. As a result, no information about the particular way in which the analysis happens to be inappropriate can be used to help produce the correct analysis. In sum, the cost of minimizing the bandwidth of communication between syntactic processing and other processing in this case is that the language understander as a whole is reduced to the crudest and most expensive possible method for producing an appropriate interpretation of the input, namely, arbitrary choice and backtracking.

Perhaps most troubling, however, is not the fact that Marcus's theory avoids the problems posed by genuine structural ambiguity, but the manner in which it does so. Simply putting such problems aside -- shifting the burden of non-determinism to the understanding process as a whole, for example -- places this entire approach on a slippery slope. One can, obviously, continue to put aside each problem that seems to require sacrificing either determinism or modularity, and narrow the scope of the theory further and further -- what has been dubbed, by one of my colleagues, "the incredible shrinking module." But, just as with the use of non-determinism, this leads ultimately to theories which are irrefutable. It is tautological that the subset of syntactic decisions that can be resolved deterministically, using only syntactic information, can in fact be resolved within a deterministic, modular theory of syntactic analysis. If the claim of modularity is to have any empirical force, therefore, it is necessary to show that the set of such decisions, and the rules needed to resolve them, can be characterized *in advance*. And it seems to me that the fact that exactly the same potential ambiguity can, in some cases, be resolved within a sentence on syntactic grounds, and in others leads to genuine structural ambiguity, immediately refutes this possibility in the case of most, if not all, syntactic decisions.

This point also bears on one of the most common arguments made in favor of modular syntactic analysis. There is a common-sense grain of truth at the root of the idea, and it is often put as follows: "What's the problem here? You let syntax do what it can deterministically, and then semantics takes care of the rest." Although even this claim is arguable -- I believe that there are many cases in which understanding need not in principle, and does not in fact in human beings, proceed by "letting syntax do what it can" -- let's suppose that it were true. It simply does not follow that there must exist an independent syntactic module which makes those particular decisions that can in fact be made deterministically using only syntactic information. It is perfectly compatible with an integrated view that decisions that don't happen,

in a particular instance, to require the use of semantics be made, in that instance, without such information -- although it need not necessarily be the case that they are. In order to justify the independent existence of a syntactic module, therefore, it is not enough that particular *instances* of syntactic decisions can be made deterministically using only syntactic information. What is necessary is that *classes* of such decisions, and the rules which are necessary to make them, be characterized and specified in advance as the proper domain of such a module.

5.2. Syntactic representations and determinism

A critical examination of how Marcus's theory addresses -- or rather, fails to address -- the need to functionally justify the output of a deterministic, modular syntactic analyzer, reveals numerous further examples of the sorts of theoretical deficiencies described above. Consider the problems raised in determining, and representing, the correct role of a prepositional phrase within a sentence -- one of the most common instances of genuine structural ambiguity in English, and moreover a problem that is widely acknowledged, even by Marcus, to require heavy use of semantic and contextual information (see also Woods, 1973). Because a syntactic analyzer is by itself incapable of correctly determining where to attach a prepositional phrase, Marcus's parser simply attaches all such phrases to the closest available constituent that is syntactically acceptable, regardless of whether or not that is correct (personal communication). For example, it would analyze the sentence "I kissed the girl in the park on the lips," as if the prepositional phrase "on the lips" modified the noun "park." The main point here is that a semantic interpreter using the output of Marcus's parser must beware of taking it too seriously -- it is likely to be incorrect. This is hardly the hallmark of a functionally constrained output. The only reason for this approach, indeed, is to avoid facing the need to use non-syntactic information to make the decision.

More recently, Marcus, Hindle, and Fleck (1983) have attempted to face up to this sort of problem by proposing that the output of a syntactic analyzer should not, in fact, be a phrase marker of the sort employed in linguistics. Instead, they propose that it should be something vaguer and less informative, a description which corresponds to a set of such phrase markers. In particular, the relation "immediately dominates" in the structural description of a sentence is replaced by the less informative relation "dominates," and constituents are referred to by non-unique names, so that two symbols may -- if the facts that are known about the constituents to which they refer are compatible -- turn out to refer to the same constituent. However, once again no functional justification is offered for this sort of output, other than the fact that it may prove possible to produce it without the use of semantic or contextual information, and once again this failure places the entire approach on a slippery slope. As long

as the output of the syntactic module can be redefined to be less informative -- without being subject to any constraints of functional utility -- whenever the attempt to produce a more informative output appears to threaten either modularity or non-determinism, then of course it will always be possible to maintain a non-integrated view. In other words, if the output of a syntactic analyzer is defined as that structural information about a sentence which can be deterministically derived without appeal to semantic or contextual information, then it is once again tautological that deterministic, modular syntactic analysis is possible. But unless it can be demonstrated that such an output will be functionally useful, the claim of modularity is without empirical content. Just as much as permitting arbitrary choice and back-up, defining the output in this way results in modular theories which are irrefutable. To take this to its logical conclusion, one can quite simply write a deterministic parser that does not require semantics: It need only take in strings and output them without alteration. Without applying the constraints of functional utility, we have no guarantee that the output of a process is anything other than a trivial transformation of the input.

Moreover, although a functionally justified output is a *necessary* condition for an empirically significant claim of modularity, it is by no means *sufficient*. It must also be shown that the information contained in such an output can actually be exploited without violating the original claim of syntactic modularity, and this is not a foregone conclusion: The utilization of such a representation may itself turn out to require the highly integrated application of syntactic, semantic, and contextual information. That is, even if it were possible to produce useful syntactic representations in a deterministic, modular fashion, such representations might prove too impoverished to support the application of useful -- or even necessary -- syntactic rules. Indeed, it is quite possible that a given syntactic rule could sometimes be applied by the syntactic module -- given a simple and unambiguous enough input -- while at other times the information necessary to apply the same rule would not be available. If such rules are genuinely useful, then it will be necessary to apply them after sufficient information about the structure of the input has been recovered -- which is to say, after semantic and contextual information have been applied. In other words, even if it proves possible to produce functionally useful, if somewhat impoverished, structural representations of input sentences in a modular and deterministic fashion, it may nevertheless be the case that actually *using* such representations in understanding entails the subsequent application of syntactic and semantic rules in a highly integrated fashion. There seems little point to the claim of modularity under such circumstances.

It is important to understand that the above argument is not merely theoretical. The output produced according to Marcus, Hindle, and Fleck (1983) is in fact insufficiently

informative to support the use of the putative syntactic rules constraining pronoun reference cited in Marcus (1984) as arguing for the need to compute explicit, independent syntactic representations. Such rules depend on knowing, rather precisely, how high in the structural description of a sentence a noun phrase is with respect to a potentially co-referent noun phrase, and this is exactly the sort of information that has been discarded by Marcus *et al.*'s later theory. For example, in the sentence "I recognized the spirit in him by the boy's behavior," determining that "him" and "the boy" can be co-referential according to these rules depends on knowing that the prepositional phrase "in him" is attached to the noun phrase "the spirit," while the prepositional phrase "by the boy's behavior" is attached to the verb phrase "recognized." However, since prepositional phrase attachment depends on semantic and contextual information, on Marcus *et al.*'s account this determination would not be made by the syntactic module. Thus, if such syntactic rules for pronoun reference were in fact to be applied in understanding, that could only occur *after* semantic and contextual information had been employed to recover sufficiently explicit information about the structure of the input utterance. Indeed, on this account, such syntactic rules would not even seem to be within the province of the syntactic module itself.

Now the fact is that I don't believe that these rules are completely correct, or that a purely syntactic account of the phenomena in question is actually possible. (A convincing critique of such claims can, in any event, be found in Bolinger, 1979.) Still, I am not sure that I would go so far as to say that syntax plays *no* role in the matter, and unless Marcus is prepared to say that that is the case, he must concede that syntactic knowledge, and rather sophisticated syntactic knowledge at that, must exist and be applied outside of the syntactic module. It is not clear, under these circumstances, what the claim of syntactic modularity amounts to.

5.3. Garden path sentences and modular parsing

One of Marcus's most ingenious arguments in favor of the determinism hypothesis was the use of garden path sentences as "the exception that proves the rule." A garden path sentence, to review, is a sentence about which an understander makes an erroneous decision during understanding and becomes consciously aware of that fact. For example, the classic "The horse raced past the barn fell," is a garden path sentence because the reduced relative subclause "raced past the barn" — "reduced" because the words "that was" have not been uttered — is originally taken to be the main predicate of the sentence, as would be correct if the word "fell" were absent.

The claim that language analysis does not entail backtracking would seem, on the face of

it, to be refuted by such examples. However, Marcus turned this phenomenon to his advantage by pointing out how difficult it is to explain within the framework of a non-deterministic approach to language analysis. He argued that there is a sense in which a language analyzer that routinely employs backtracking is constantly being led down the garden path to incorrect analyses, even during the analysis of inputs which do not, to humans, seem to be problematic. How then could such a process distinguish between the back-up arising in garden path sentences -- which is conscious -- and the unconscious back-up routinely employed in analyzing unproblematic inputs? In Marcus's own theory, on the other hand, the difference is easily accounted for: Garden path sentences are exactly those in which the usual deterministic language analysis mechanism fails.

However, it is possible to agree with Marcus that garden path sentences argue against the indiscriminate use of backtracking in language understanding, without agreeing that they constitute a purely syntactic phenomenon, and that they therefore provide evidence for his theory of modular deterministic syntactic analysis. Indeed, in view of Crain and Steedman's (1985) experimental results, this is the only interpretation possible. Crain and Steedman demonstrated that the phenomenon of garden path sentences is crucially dependent on semantic and pragmatic factors, and, therefore, that no purely syntactic account -- including Marcus's -- can be correct.

Consider their example "The children taught by the Berlitz method passed." Although this sentence is syntactically identical to "The horse raced past the barn fell," it is not a garden path sentence. That is, since it is unlikely that children teach, but quite likely that they are taught, the phrase "taught by the Berlitz method" is assumed to be a reduced relative subclause describing which children are being referred to. Crain and Steedman explicitly contrasted this sentence with "The teachers taught by the Berlitz method passed," which -- although differing only in that the word "teachers" has been substituted for the word "students" -- clearly is a garden path sentence. These examples demonstrate, as Crain and Steedman pointed out, that the phenomenon of garden path sentences can only be explained by assuming that factors of semantic and contextual plausibility play a role in syntactic decisions during sentence analysis. In sum, they showed not only that this phenomenon cannot be explained within the framework of Marcus's theory, but that it actually constitutes strong evidence against that theory.

One final comment. Marcus's theory of garden path sentences rests on the claim that people do not backtrack and produce an alternate analysis for an input without becoming consciously aware of that fact. However, since his account of genuinely ambiguous sentences entails the use of such backtracking as a matter of course, the question naturally arises of why

people do not seem to be conscious of that fact in such cases. I suspect that Marcus would be tempted to respond that, in the case of genuine structural ambiguity, the discovery of the error and the subsequent decision to back up and produce another analysis would not reside in the syntactic module itself, as it does in genuine garden path sentences, and that this explains the difference. If he did so, however, he would be forced to accept the conclusion that the garden path sentences devised by Crain and Steedman really do reflect erroneous *syntactic* decisions by the putatively independent syntactic module, despite the fact that those decisions are demonstrably made on the basis of semantic and contextual information. In other words, Marcus's account of garden path sentences is not merely empirically incorrect; it appears to be inconsistent with his account of genuinely ambiguous sentences.

5.4. Lexical ambiguity, modularity, and determinism

One of the most important and difficult problems in language analysis is the timely resolution of lexical ambiguity. For the understanding process as a whole, the problem is one of determining the appropriate meaning of an ambiguous word. It is widely understood -- although, as we will see in the next chapter, just as widely ignored -- that such disambiguation in general requires the heavy use of semantic and inferential processing. For a syntactic analyzer, however, lexical ambiguity manifests itself in the somewhat narrower problem of determining the appropriate part of speech of a word that belongs to several syntactic categories. Nevertheless, these two problems are related: Semantic ambiguity often -- although not always -- entails part of speech ambiguity. The reason why this poses a problem for syntactic analysis, of course, is that part of speech ambiguity is one of the chief causes of structural ambiguity in natural languages. Since much of the structural ambiguity of language arises as a result of lexical ambiguity, and since the resolution of lexical ambiguity seems, on the face of it, to require heavy use of semantics and inference, this problem constitutes one of the primary motivations for an integrated approach to language analysis.

For the very same reasons, lexical ambiguity would appear to be a crucial issue for any theory that purports to show how syntactic structural ambiguity can be resolved in a deterministic and modular fashion -- that is to say, with limited look-ahead and highly restricted use of semantic and contextual information. Nevertheless, lexical ambiguity is another of the problems that Marcus's theory simply puts aside: With one or two exceptions, words are taken to be syntactically unambiguous in his work. Once again, however, this strategem raises the question of exactly what empirical claims are being made by the theory. At the risk of belaboring this point, it is undoubtedly true that by ignoring all problems that seem to require sacrificing either modularity or determinism, one can construct a modular, deterministic

syntactic analyzer. But unless it can be shown that the output of such a parser would be useful, and that its use would not itself entail the joint application of syntactic and semantic knowledge, such a claim is without empirical import.

Moreover, lexical ambiguity is precisely the sort of problem that militates against the possibility that those conditions hold true. Indeed, the one or two cases of lexical ambiguity that Marcus does attempt to resolve within the framework of his theory simply serve to show how profound the impact of the problem actually is. For example, in order to disambiguate whether the word "have" is used as an auxiliary or a main verb, Marcus introduces a diagnostic rule which is arguably the most complex in his entire grammar. Nevertheless, as Marcus himself points out, the rule fails on many obvious examples. How well such rules would work in the context of many *other* ambiguous words is highly questionable. Indeed, Milne's (1982) attempt to address lexical ambiguity within the framework of Marcus's theory led to a substantially greater reliance on semantics. One need not agree with the details of his proposals to find this result suggestive.

6. Conclusion

In this chapter, we have seen that current AI theories of modular syntactic analysis are not well justfied functionally. They are, in a sense, "process models," but it is not clear that they are models of any real process. The nearly complete absence of adequate functional constraint on the representations which constitute their output has led, inevitably, to a scientifically fatal lack of constraint on the processes which are proposed to produce that output. As long as outputs are defined and redefined at the whim of the theorist, without being functionally justified, then of course any process at all can be "proven" viable. The ultimate role of such a process within a functional theory of language understanding, however, remains in doubt.

Indeed, the most popular models of language analysis -- ATNs and Prolog-based parsers -- are not process models at all, but simply programming languages. As I argued in chapter one, the claim of modularity is irrefutable if the indiscriminate use of arbitrary choice and backtracking is permitted. Such models are, therefore, in principle incapable of either asserting or supporting an empirically meaningful claim of syntactic modularity. Marcus's more recent deterministic theory of syntactic analysis seems at first glance more promising. However, by its repeated failure to address many of the problems which make language analysis so difficult in the first place -- such as lexical ambiguity and genuine structural ambiguity -- this theory too fails to assert or support any empirically significant claim of syntactic modularity. Moreover,

the very problems in language analysis that this theory ignores provide strong evidence that such claims are in fact insupportable.

From the functional perspective that motivates an integrated approach to language processing, the theoretical shortcomings of modular approaches to syntactic analysis are not really very surprising. After all, the roots of such theories lie not in a functional view of language as a goal-directed behavior, but rather in the descriptive view of language behavior taken by linguistics. However appropriate such a view may be for linguistics -- and it is by no means a universally accepted perspective even within that field -- it is entirely inappropriate for AI. The decision as to whether a given process should or should not employ a given class of information in the performance of its task must be made on functional grounds, weighing the utility of the information against the cost of gathering and applying it, rather than on *a priori* descriptive grounds. Similarly, the task itself must be defined functionally, in terms of its utility to the organism, rather than by its conformance to some *a priori* descriptive framework. One can argue for a modular approach to language analysis, and against an integrated approach, but the functional premisses that motivate an integrated approach provide the only basis upon which such an argument can be fruitfully carried out. The paucity of the results stemming from modular approaches to language analysis, despite the investment of enormous scientific effort, directly reflects the failure to keep such functional considerations foremost in mind.

CHAPTER 4

LEXICAL AMBIGUITY AND VAGUENESS IN LANGUAGE ANALYSIS

1. Introduction

Lexical ambiguity is one of the most basic, yet problematic, characteristics of natural language. It is, first of all, far more pervasive than it intuitively appears to be: Because people are not consciously aware of most of the ambiguities in what they read or hear, the fact that most of what they read or hear *is* ambiguous is not immediately apparent. However, a glance at any ordinary dictionary should make it plain that lexical ambiguity is extremely common.

The importance of lexical ambiguity, and the difficulties involved in attempting to resolve the problem, have been apparent since the original work in machine translation thirty years ago. Bar-Hillel (1960), in his critique of that work, showed that determining the correct sense of an ambiguous word depends, in general, on plausible inferences from extremely complex features of the context in conjunction with arbitrary facts about the world. He gave as an example the problem of choosing the correct meaning of the word "pen" in the sentence "The box is in the pen." In this sentence, the pen in question is probably an enclosure, such as a play-pen, rather than a writing implement. Bar-Hillel argued that in order to determine this, a language analyzer would need access to background knowledge of the functions and relative sizes of these two different kinds of objects, and to information about the context in which the sentence appeared, as well as some means of using that knowledge to determine which of the various possible interpretations of the sentence was the most plausible.

Of course, lexical ambiguity is not just a problem for *semantic* analysis. As I pointed out in the last chapter, it is also one of the chief causes of structural ambiguity, and, therefore, an issue with which syntactic analyzers must contend as well. This aspect of the problem has also long been appreciated. In the well-known example "Time flies like an arrow," (Kuno, 1965), much of the structural ambiguity of the sentence stems from the part-of-speech ambiguity of the words "time," "flies," and "like," which in turn reflects their semantic ambiguity.

Thus, on the one hand, the resolution of lexical ambiguity seems to require complex plausible inference based on contextual information as well as world knowledge; and, on the other, it has a profound impact on syntactic processing, since it is a chief cause of structural ambiguity. As I argued in chapter one, in order to avoid arbitrary choice and the undirected search that inevitably results, such ambiguities should be resolved as early as possible, and on the basis of as much information as possible. Lexical ambiguity is, therefore, one of the characteristics of natural language that argue for an integrated approach to language analysis, one in which the inferential and memory processing that is required for disambiguation is intimately involved in the analysis process itself.

2. Lexical ambiguity and syntactic analysis

In syntactic analysis, the problem of lexical ambiguity is not one of choosing the correct sense of a word, but simply the correct part of speech. However, as Kuno's (1965) well-known example demonstrates, these problems are closely related. Word-sense ambiguity very often entails part-of-speech ambiguity as well, and therefore correctly disambiguating the part of speech of a word will in general depend on complex semantic and pragmatic processing. As a result, syntactic analyzers cannot be expected to solve by themselves the problem of lexical ambiguity, even just part of speech ambiguity. Nevertheless, it is not unreasonable to expect that they might contribute to its solution.

However, as we saw in the last chapter, the chief approach to resolving syntactic ambiguity, lexical or otherwise, is simply to try each alternative, while being prepared to back up in case it should prove mistaken. This is the approach taken in ATN parsers and descendant models (see, e.g., Thorne, Bratley, and Dewar, 1968; Bobrow and Fraser, 1969; Woods,1970; Colmerauer, 1978; Pereira and Warren, 1980). When such a parser encounters an ambiguous word, it simply tries each possible choice for that word's part of speech which will enable a transition, and which therefore offers the possibility of successfully parsing the input sentence according to the grammar utilized. If the choice does not lead to a successful syntactic analysis, then it will be discarded when the parser backs up. This process will be repeated for a given word each time the parser encounters it when traversing the network. All and only the choices that lead to successful analyses will be retained as the parse's analyses. Further disambiguation is the responsibility of the semantic and pragmatic components of the understanding process.

The general critique of arbitrary choice and backtracking presented in the last chapter applies equally well to the specific application of such techniques to

disambiguation. Nevertheless, there is currently no other theory of lexical disambiguation in syntactic analysis: Despite the fact that lexical ambiguity is a prime cause of much of the structural ambiguity in language, Marcus's more recent deterministic approach to syntactic analysis basically ignores the problem, and loses much of its empirical significance as a result. In sum, no adequate syntactic theory of lexical disambiguation exists. This is not particulary surprising, since lexical ambiguity is, after all, primarily a question of word-sense ambiguity, rather than just part of speech ambiguity, and so primarily a semantic problem rather than a syntactic one. Despite the problems it raises for syntactic analysis, syntax alone cannot be expected to do very much about lexical ambiguity.

3. Lexical ambiguity and semantic analysis

Quite naturally, the issue of lexical ambiguity has received far more attention in theories concerned with semantic analysis than in those concerned simply with syntactic analysis. However, although at first glance there seem to be a variety of different semantic approaches to 'lexical' disambiguation, most approaches in fact turn out to share only one or two fundamental mechanisms. By far the most widespread of these methods is the use of *selectional restrictions* (Katz and Fodor, 1963). These are semantic requirements associated with the structures representing the meanings of words or phrases, which must be met by another semantic structure before the two can be combined. For example, an action like eating might require that [illegible] be animate. In general, selectional restrictions are one-place predicates that test for the presence or absence of some semantic feature, or some boolean combination of such [illegible]

[illegible] restrictions in disambiguation is, in principle at least, quite [illegible] chooses the sense (or senses) of a word that selectional [illegible] other semantic structures in the sentence, either because [illegible] structures, or because they fulfill the requirements [illegible] example from Katz and Fodor (1963), consider [illegible] mean, among other things, either a fancy party with [illegible] In the sentence "John hit the ball," the use of [illegible] choosing the round object sense of "ball," since the [illegible] object but not to a social gathering. Of course, if a [illegible] were available, or could be generated as a novel [illegible] to disambiguate either "ball" or "hit" in this [illegible] The fact that even such a simple example poses [illegible] something about its general utility. Indeed, in view

of its popularity, it is rather striking how difficult it is to produce examples of lexical ambiguity which can be resolved unequivocally by the use of selectional restrictions.

A variety of different methods have been developed for applying selectional restrictions in the resolution of lexical ambiguity; I will briefly mention just a few of them here. Winograd (1972) proposed that they be used by semantic interpretation specialists associated with functional syntactic constituents such as noun groups and clauses. Riesbeck (1975) proposed encoding selectional restrictions in the tests of the lexically indexed productions that represent, in his theory, the different meanings of a word. Rieger and Small's (1979) theory of word experts and Hirst and Charniak's (1982) theory of Polaroid words are based on more sophisticated versions of this idea. Wilks (1976) has proposed that selectional restrictions should not be absolute requirements, but simply preferences. In his model, one picks the sense of each word that maximizes the total number of preferences satisfied in a given sentence.

The other major approach to handling lexical ambiguity, one that attempts to take into account the extra-sentential context of an ambiguous word, involves the use of a *scriptal lexicon* (Schank and Abelson, 1977; Cullingford, 1978; Riesbeck and Schank, 1978; Charniak, 1981). This idea is based on the observation that many words have special meanings in particular contexts. Thus, in a sense, each script or frame used in understanding a text should have an associated lexicon in which words are assigned their script-specific meaning. For example, the script for a baseball game would have an associated lexicon in which the word "home" would be defined as the plate in the ground over which batters stand, and which a player must touch to score a run. By itself, this idea is not very useful for disambiguation, except insofar as it keeps script-specific meanings out of consideration unless the relevant script is "active." The crucial simplifying assumption which is usually made, therefore, is that if a given script is "active," all words in its scriptal lexicon can be presumed to have their script-specific meaning. In other words, if the context for baseball game were active, then the word "home" would simply be assumed to mean "home plate." No other sense of the word "home" would be considered.

Although both selectional restrictions and scriptal lexicons are very useful up to a point, they both have severely limited applications, it should be clear that they have severe limitations. The assumption which underlies the scriptal lexicon approach, that words will not appear in their script specific sense, is clearly not true. For example, consider the following context of a story about a baseball game: "The game was so lopsided that we left home after the seventh inning." Here, the home in question is

probably Fred's residence, not home plate.

The use of selectional restrictions has similar limitations. Consider the following variant of Bar-Hillel's example: "The pen is in the box along with assembly instructions." Here, the pen in question is probably a play-pen, and almost certainly not a writing implement. Determining this requires recognizing that the assembly instructions are probably for the assembly of the pen, and knowing that play-pens often require assembly by the consumer after purchase, whereas writing implements do not. Using this knowledge in turn requires inferring that since the pen is in a box with assembly instructions, it has probably just been purchased by the consumer. The point here is that these are simply not the kinds of rules that can be represented and employed as selectional restrictions, except at the risk of precluding the correct analysis of other examples. We cannot, for example, just invent a feature "objects that can be assembled" as a selectional restriction on the object of "assemble," and which would be a property of play-pens but not writing implements. Writing pens certainly *are* assembled, in factories, and they may even be assembled by the consumer, as in "John assembled the pen after cleaning it and putting in a new cartridge."

4. Lexical ambiguity and integrated analysis

The above discussion makes it clear that what must be brought to bear on the problem of lexical ambiguity are the general inference and memory processes used in understanding. Thus, lexical ambiguity is one of the problems in language analysis which motivates an integrated approach, one in which inference and memory processing play an important role in the analysis process itself (Schank, Lebowitz, and Birnbaum 1980, Schank and Birnbaum 1984). It is therefore crucial that integrated theories address the issue of how inference and memory can be applied in the early resolution of lexical ambiguity. The failure to do so would simply undercut the rationale for pursuing an integrated approach in the first place. Despite its importance, however, surprisingly little attention has been devoted to the issue.

For example, consider the approach taken in the model proposed by Dyer [illegible] explicitly aims to be a model in which memory [illegible] language analysis process. Despite [illegible] is limited to the use of [illegible] the tests of lexical [illegible] best see how the [illegible] process [illegible]

If the actor is a VEHICLE or
the SCENARIO is TRANSITIONAL with a VEHICLE instrument,
Then interpret "run into" as VEHICLE-ACCIDENT;
Else If the object is a HUMAN who has an
INTERPERSONAL RELATIONSHIP with the actor,
Then interpret "run into" as RENEW-INTERPERSONAL-RELATIONSHIP.

Let's analyze how this procedure is intended to work. The test for whether or not the actor is a vehicle is simply a selectional restriction. The test for whether "the SCENARIO is TRANSITIONAL with a VEHICLE instrument" is perhaps more puzzling. However, its purpose would seem to be to handle examples such as "While I was driving home, I ran into a parked car," in which the actor of "run into" does not appear to be a vehicle, but the proper interpretation is nevertheless vehicle accident. In effect, this is simply an implementation of the scriptal lexicon idea: If the vehicle travel frame is active, then "run into" means vehicle accident. Both of these rules are subject to the limitations described in the last section. For example, this use of the scriptal lexicon approach would fail on the following text:

While I was driving home, I remembered I needed some milk. I ran into a Seven Eleven and picked up a half-gallon.

Finally, let's consider the test for a human who has some interpersonal relationship with the actor. Here, the model begins to employ knowledge beyond simple selectional restrictions, which are technically just one place predicates. The problem is, it still *employs* this knowledge exactly as if it were just a selectional restriction. Although the presence of such a relationship — or, in fact, any semantic feature — is indeed the sort of knowledge that may be relevant in ascertaining the correct meaning of a word, its use as a sufficient condition in a non-inferential, lexically indexed rule of this variety is entirely misplaced. The point is that such knowledge must be represented and indexed in a way that makes it available for use by the general [illegible] capabilities of the understander.

[illegible]

would exhibit such behavior, and thus enable the understander to construct a causally coherent representation of a textual fragment describing such an episode. It is the attempt to construct such a causally coherent representation that determines the proper interpretation of "run into." A particular interpretation, such as "social encounter," is preferred to the extent that it promotes such coherence.

But the rule cited above does not explicitly represent such causal knowledge, nor does its choice of an interpretation for "run into" depend on the attempt to infer a causally coherent representation. Instead, the inference process is "short circuited" by directly linking some (but not all) of the relevant features with some (but not all) of the possible interpretations. Such a rule simply cannot work in general. Consider, for example, the following text:

> John was racing down the street trying to catch a bus. All of a sudden, his neighbor Fred stepped out of a doorway into his path. John ran into Fred and knocked him down. Fortunately, he wasn't hurt.

What both this example and the previous one demonstrate, to repeat, is that the proper interpretation of "run into" should be determined on the basis of the attempt by inferential memory to construct a causally coherent representation of the text as a whole -- which is, after all, one of the chief functions of inferential memory in understanding. In a genuinely integrated model of language understanding, it must be on the basis of these sorts of inferential considerations that language analysis problems, such as lexical ambiguity, are resolved. Instead, in Dyer's model we find that such inferential processing occurs only *after* a word has already been disambiguated by means of selectional restrictions and scriptal lexicons. Inferential memory can therefore play no role in the disambiguation of a lexical item -- or indeed, as far as I can determine, in the solution of any other problem in language analysis.

The model of integrated partial parsing proposed by Schank, Lebowitz, and Birnbaum (1980) and substantially extended by Lebowitz (1980) also depends, primarily, on the scriptal lexicon approach -- in fact, most words are simply as ambiguous as far as the model is concerned, since it presumes that input stories will involve only a single domain. In some ways, however, this model constitutes a much more serious attempt to use inferential memory in disambiguation. In order to construct coherent representations of input stories, the model employs a version of script application (Schank and Abelson, 1977; Cullingford, 1978) -- in which an action or state is explained or interpreted by matching a scriptal expectation. The model can then use these expectations to disambiguate a word by choosing the meaning that satisfies one of them. This method was originally proposed by Riesbeck and Schank (1978).

This method is clearly a step in the right direction. However, its successful application depends on two fairly restrictive conditions. First, it assumes that if a script is active, then an ambiguous word must have the meaning that matches an expectation from that script. Second, it can only work in those cases when exactly one meaning of a word matches a scriptal expectation. But consider what is likely to happen when more than one script is active, or when the scripts are larger and more detailed, or when expectations from sources other than scripts are utilized in understanding -- in other words, under the conditions that usually hold in language understanding. Under such conditions, it seems almost certain that more than one meaning of an ambiguous word would match an expectation, or to put this another way, that more than one interpretation could appear, on the basis of such a simple matching process, to be coherent within the context.

Thus, this method for disambiguating on the basis of criptal expectations in will not work in many situations: It will either fail to disambiguate, or else simply choose in a way which guarantees a high probability of error. The method can only be employed reliably when only one script is active, and when only one sense of the word matches an expectation from that script. In other words, this use of scriptal expectations is virtually equivalent, in the power and scope of its disambiguation capabilities, to the use of a scriptal lexicon: As far as disambiguation is concerned, one might just as well explicitly stipulate that the given word will have a given meaning if the given script is active. The main advantage of the approach, then, resides in the fact that such stipulations need not be explicitly given.

How can the the use of scriptal expectations, or more generally of contextual information from varying sources, be extended to handle those cases in which more than one meaning of an ambiguous word might seem at first to fit the context? Several factors must be taken into account beyond the mere occurrence of a match between a potential word meaning and an expectation. First, which expectations are more important, or more likely to be satisfied at this point in the text? To put this in more general terms, which of the explanations for the different possible interpretations is more plausible or more salient? Second, does the text supply any additional clues? For example, a candidate semantic structure may be the right sort of action to satisfy an expectation, but may nevertheless be inappropriate because its potential actor, as specified in the text, does not match the binding already assigned to the actor in that expectation. The use of such information is essential to exploit the full potential of memory and inference in lexical disambiguation.

In fact, this last requirement poses the greatest challenge to recent models of language analysis employing *connectionist mechanisms* (see e.g., Small, Cottrell, and Shastri, 1982),

Cottrell, 1984; Waltz and Pollack, 1984; for an earlier attempt, see Quillian, 1968). The manipulation of variables and variable bindings is a difficult issue in the connectionist framework (J. Feldman, personal communication), and as currently formulated these models do not seem capable of utilizing such information in disambiguation (see also Minsky, 1968, for a discussion of the problem of variable binding in parallel computation). Thus, they are subject to the exactly the same limitations as described above -- they appear to be, in effect, simply a rather novel approach to the implementation of scriptal lexicons. Whether the clever manipulation of parameters such as weights and activation levels can overcome these limitations remains to be seen. One possible solution is to use connectionist methods simply to suggest potential inference chains, and employ more traditional inference mechanisms, capable of manipulating variable bindings, to check over the suggestions (Charniak, 1983). Another possibility, requiring a more radical change in the connectionist framework, is to allow variable bindings to be passed between the units in a memory network (Riesbeck and Martin, 1986).

5. Vagueness and ambiguity

In chapter two, I pointed out the importance of distinguishing between true lexical ambiguity and vagueness. This distinction is particularly important in light of the need for inference to determine the appropriate interpretation of some word or phrase in some context, as argued above. True ambiguity arises when a lexical item points directly to more than one conceptual structure in memory, each of which represents a separate and distinct meaning for that word. The resolution of lexical ambiguity is thus a problem of *selecting* the appropriate meaning from among a specified set of potential meanings for a lexical item. Vagueness, in contrast, arises when a word or phrase -- e.g., the word "get" -- points to only one conceptual structure in memory, representing a single somewhat general meaning which must be further specified on the basis of context. Unlike lexical disambiguation, therefore, the resolution of vagueness cannot be construed as a problem of selection from among a pre-computed set of potential interpretations, but rather entails the active *construction* of an appropriate and sufficiently specific interpretation for an input on the basis of context. Rather than simply being retrieved on the basis of a few clues, in other words, the specific interpretation that is appropriate must, in general, be derived on the basis of inference. As a result, non-inferential techniques such as the use of selectional restrictions or scriptal lexicons -- whatever their merits or deficiencies in lexical disambiguation -- are in principle inadequate for the resolution of vagueness, since they fundamentally construe the problem of determining the appropriate interpretation for a lexical item as a selection problem.

It is quite likely, in fact, that this fundamental limitation lies behind the widespread

failure to distinguish between vagueness and ambiguity in the first place. If a vague word is treated as if it were ambiguous, then such non-inferential techniques as the use of selectional restrictions can be applied, at least under certain restricted circumstances, to determine its appropriate interpretation; otherwise, they cannot be applied at all. In other words, if the problem of constructing an appropriately specific interpretation for some word or phrase in context is transformed into a problem of selecting an appropriate interpretation from among a pre-computed set of potential interpretations, then that selection can be based, in a limited domain at least, merely on the presence or absence of certain contextual features. The problem of inference can be pretty much ignored.

Ultimately, however, an approach based on the idea that meanings are selected rather than constructed cannot work except in a fixed, limited domain. Human language use is notable for its flexibility and extensibility, for the ability to say new things using old words and phrases. It is, indeed, becoming increasingly clear that such phenomena as metaphor are central elements of language and thought, not peripheral issues to be dealt with once the central problems have been solved (see, e.g., Lakoff and Johnson, 1980; Carbonell, 1982). Thus, for example, it is quite easy for us to generate and understand slightly novel uses of a word. In a model which determines the appropriate interpretation of a word simply by selecting from among a pre computed set of possible meanings, however, correctly understanding even slightly novel uses of a word would be out of the question.

Consider, for example, the phrase "run into," which we discussed in the last section. This phrase poses problems quite similar to those raised by such a word as "set." The phrase seems to mean different things on different occasions — indeed, it seems to mean so many different things that to assume that it is [illegible] to the [illegible] that it has scores of distinct senses. Yet these [illegible] typically the case with [illegible] the phrase [illegible]

[illegible]

When I rode the bus home from work today, we ran into an old man selling flowers.

When I rode the bus home from work today, I ran into an old man selling flowers.

We ran into some good luck.

We ran into a tough problem.

I ran into a great little wine.

I turned left, went down the hill, and ran into Whitney Avenue.

What most of the above utterances seem to have in common is the idea of a sudden, most likely unintentional, and possibly unexpected, encounter with some thing or some situation, and further, that the encounter in some way affects the continuation of whatever activity the agent is engaged in at the time. This, I maintain, is all that "run into" by itself really means. This is what you would understand it to mean if you ran into the sentence "I ran into a bloop with blap the other day," not knowing what "bloop" or "blap" were. In other words, "run into" is primarily an extremely abstract phrase, not an ambiguous one.

On the other hand, all of the above sentences do in fact mean different things. Some are best interpreted as describing a vehicle accident, others a social encounter, others a barrier to the solution of some problem, and still others the unexpected discovery of something. Some of these meanings, because of repeated prior usage, might be pre-computed and stored in memory. But even if this is the case, determining which meaning is most appropriate will require inference, and that is the key to resolving vagueness as well. The phrase "run into" need not mean "vehicle accident," or "social encounter," or "discovery," even though the sentences that contain it might mean those things. The only way to resolve this paradox is to utilize inference to construct the appropriate specific meaning, taking into account both the rather abstract meaning of "run into" and contextual information.

Such an approach requires addressing three issues. First, it requires a way to represent the abstract concepts, such as "encounter," "sudden," and "unintentional," in terms of which the abstract meaning of "run into" could be represented. Second, it requires a context in which specific correlates of such abstract concepts are represented. And third, it requires an inference process capable of sorting out which potential specific correlates of abstract concepts might be intended in an utterance — in other words, which would lead to the best joint fit between utterance and context.

For example, let's consider how the meaning of "While driving home from work, I ran into a traffic jam." might be derived inferentially. "Run into," by assumption, simply means a

sudden, and probably unintentional, encounter with something. Most of the information about what is going on here, then, must come from "driving" and "traffic jam." It seems quite plausible that the driving script makes note of possible problems that might arise in the course of driving somewhere -- e.g., accidents, mechanical failure, heavy traffic, and so on -- and which, if encountered, can can impede progress towards the goal of getting somewhere. Since they impede progress towards the goal of the script, it can be inferred that any encounter with them is most likely unintentional. On an admittedly somewhat simplistic analysis, then, we can see that this inference permits a "match" between the input -- a sudden and unintentional encounter with heavy traffic -- and the known problem of encountering heavy traffic when driving -- a type of encounter which, because it is a problem, is presumably unintentional. The output that would hopefully result from such an inference process would be an instantiation of the "caught in traffic" memory structure, which might be paraphrased something as follows: "While driving home from work, I suddenly and unintentionally found my car in the midst of a large number of other vehicles, travelling quite slowly, for some distance, which forced me to drive my car very slowly as well."

Although most of the above interpretation is derived from information in the driving script, there is still something contributed by "run into," namely, the notion of "sudden and unintentional." Thus, for example, "While driving home from work, I ran into heavy traffic," seems to mean something a bit different from "While driving home from work, traffic was heavy." In both cases, we understand that while driving, heavy traffic was encountered which impeded progress towards the destination. What "run into" seems to contribute in the first sentence is the idea that the encounter with heavy traffic was a bit sudden, which in turn implies that traffic was not necessarily heavy for the whole trip.

Let's look at another example: "While driving home from work, I ran into a parked car." Here, "I" is a person in the role of driver of the vehicle. As before, "run into" is taken to mean "sudden and (probably) unintentional encounter with something," and in this case the parked car is that something. In the driving frame, we can expect to find represented, among the problems we might encounter, sudden and unexpected physical contact with physical objects that are in the road -- namely, that is, facts about auto accidents. Since a parked car is something that could be in the road, we can infer that an accident took place, with the parked car being the object hit.

So far, we have not considered the role of any prior biases in the process of interpreting vague words. But consider the following story: "While I was driving home from work today, I had an accident. I ran into a parked car." Here, it seems clear that by the time that "run into"

is read, since the kind of sudden and unexpected encounter has already been specified -- an accident of some kind -- we are likely to conclude that "run into" may refer to this accident, and therefore to determine what kind of accident it was, even before the words "parked car" are read or heard. This actually becomes comical, with a sense of punning, given a story like "While I was driving home from work today, I had an accident. I ran into my ex-wife." These examples argue that interpretation of vague words and phrases must proceed as outlined in chapter two: If some specific situation in a frame is already active, and if a vague word we encounter might refer to an instance of that already active situation, then one can presume that it is in fact doing so. Thus, since the context of "auto accident" has already been explicitly established in the above stories, when "run into" is seen, its abstract meaning of "suddenly and unexpectedly encounter (or make contact with) something" matches the already active "auto accident" situation.

But how does this matching process work? In general, this is one of the most difficult problems that must be faced in attempting to implement this sort of scheme. But in many cases, there is a reasonably simple solution: If "auto accident" is already categorized in the driving frame as a sudden and unintentional encounter with an obstruction, in a way that is similar or identical to the representation of the general concept that "run into" points to, then the matching process will simply have to check for that identity. In fact, it will generally be the case for any plan that one can "run into problems," and this argues for the representation in memory of the more general notion that sudden and unexpected problems may arise in any plan. Thus, we can consider examples like "We were trying to buy the house, but we ran into those high interest rates," which means, since borrowing money is in general a prerequisite for buying a house, that sudden and unexpected difficulties arose in getting a mortgage because of the high interest rates, which in turn precluded buying the house.

The context imposed by discourse conventions can also be expected to play a role in the interpretation of vagueness. Consider the following example: "While I was walking home from work yesterday I ran into an 1843 Rochester Forge fire hydrant. What a beauty!" Here, it is clear that the appropriate interpretation is not one of physical collision, but rather discovery. The reason is that we cannot otherwise account for why the speaker is giving us all of the details about the historical origins of this particular fire hydrant (see, e.g., Grice, 1975). Indeed, it must be that the speaker is a buff of antique industrial artifacts, particularly fire hydrants.

The approach to vagueness that I have sketched out above views the problem as a simplified form of metaphor. That is, if we take the position that metaphor is based on abstract

commonalities between the concepts employed in a metaphorical description and the situation which they are being used to describe, vagueness can be viewed as the result of describing the situation directly in terms of those abstract common properties. On this view, then, the process of understanding what is usually taken to be literal language appears quite similar to the process of understanding metaphorical language. For example, consider the pair of sentences "Joe threw out the garbage," and "Judge Bean threw out the case." The former would generally be deemed a literal use of the phrase "throw out," while the latter would probably be considered metaphorical. In contrast, a theory of understanding based on the sort of processing described above would operate in more or less the same way given either as input, starting from an abstract, rather crude description of the meaning of "throw out," and using that in conjunction with contextual information to inferentially determine a more specific and appropriate interpretation.

Finally, I want to make it clear just how difficult the above examples actually are. The interpretation of a vague or ambiguous word is often assumed to take place within a relatively fixed surrounding context, the interpretation of which is not itself problematic. In real-world natural language, nothing could be further from the truth. Almost all of the words and phrases in the above examples are themselves ambiguous, vague, or elliptical. For example, in addition to the vagueness of *"run into," there is also the ambiguity or vagueness of the word* "drive." Even the phrase "drive home" is ambiguous, as in "to drive home a point in an argument," or "to drive home a nail." The word "work" here is an elliptical reference to "the place in which I work," rather than to the type of work. The word "traffic" is ambiguous, as in "the traffic in cocaine is lucrative," as is the word "jam," meaning either a type of food, a problem, or a blockage of some kind (the second meaning itself being a metaphorical extention of the third). Ambiguity, vagueness, metaphor, and ellipsis really *are* ubiquitous in language. They are not exceptional phenomena to be addressed once the central characteristics of language are understood. They are themselves to be counted as central characteristics, and they must be taken into account from the very start.

6. Ambiguity and explanation-based understanding

Although the necessity for an integrated approach to the resolution of lexical ambiguity should be clear from the above discussion, the difficulties inherent in carrying out such an approach should also be apparent. These difficulties reflect not so much on the state of theories of language analysis as on the processes of inferential memory. The importance of the problem of lexical ambiguity, therefore, stems from the functional constraints it imposes on the understanding process as a whole.

Chief among these is the lesson that the knowledge needed to draw inferential connections in understanding cannot be packaged in isolated rules that commit the understander to certain inferences irrespective of what other rules may propose. It is true that one possible interpretation of someone "running into" another person is as a fortuitous encounter leading to a social interaction. It is also true that one explanation for why two people would care to engage in such an interaction would be if they already knew each other. Thus, the knowledge that two people knew each other would provide support for interpreting "run into" as a social encounter, since such an interpretation would enable the understander to explain certain aspects of the situation. But, as we have seen, the decision that this is the *correct* interpretation cannot be made without considering the need to explain other aspects of the situation, aspects which may have nothing to do with social interactions and to which rules explaining such interactions cannot be expected to attend.

This last point bears particular attention. No single explanatory inference rule can be expected to attend to all the aspects of a situation which might affect the truth or relevance of the explanation it offers, and hence the validity of the interpretation it prefers for some vague or ambiguous linguistic element. Thus, determining which explanations to accept, and hence which interpretations to prefer, cannot be left to the inference rules themselves. Rather, there must be a more general *inferential mechanism* that determines which explanations to accept, taking into account the need to explain diverse aspects of a situation, and the evidence of diverse rules.

Probably the most ambitious attempt in this direction has been McDermott's (1974) model, which is capable of considering several potential explanations for a situation in parallel as it unfolds and choosing among them when evidence is available, as well as patching or replacing explanations that prove erroneous. Granger (1980) and O'Rorke (1983) propose models with this last capability as well, and Granger, Eiselt, and Holbrook (1986) have proposed a model of language understanding and lexical disambiguation which makes use of such techniques. The most salient feature of these models is that they *explicitly* employ criteria, however crude, for deciding whether an explanation is adequate, when one explanation is preferable to another, and when an explanation has gone awry. For example, McDermott's criteria are, basically, coherence -- an explanation must fit the facts -- and parsimony -- an explanation with fewer unjustified assumptions is preferred (see also Wilensky, 1983). The use of such criteria would seem to be a crucial aspect of any inferential mechanism capable of fulfilling the requirements set out above, and thus capable of resolving lexical ambiguity in a general manner.

In sum, an examination of the problem of lexical ambiguity makes it clear that the most appropriate interpretation of a word, phrase, or utterance can only be determined on the basis of the attempt to infer the best explanation for the input as a whole. Moreover, because we are able to use language to describe novel situations, or to understand such descriptions, it is impossible to determine the appropriate interpretation of words and phrases on the basis of context. In the face of these requirements, any attempt to determine the most appropriate interpretation of a word or phrase on the basis of a pre-specified and fixed set of necessary and sufficient conditions associated with that word or phrase must inevitably fail. Indeed, it seems impossible construct simple, pre-specified tests -- or, for that matter, not so simple ones -- that correctly state just the necessary conditions for an interpretation, and which could therefore be used to rule out its applicability with complete certainty. There is little chance, therefore, that it will be possible to construct a model of lexical disambiguation -- and, *a fortiori*, of the understanding process as a whole -- that works by applying a series of such simple tests in succession, gradually narrowing the space of potential meanings until an appropriate interpretation is converged upon. Only one rule must be in error to ruin the entire result. Rather, the understanding process must be flexible enough to take into account diverse sorts of information from a wide variety of sources in determining the interpretation of an input, and capable of doing so early enough to avoid the accumulation of errors. It must, in other words, be an integrated process.

CHAPTER 5

INTEGRATED UNDERSTANDING AND THE USE OF THEMATIC KNOWLEDGE

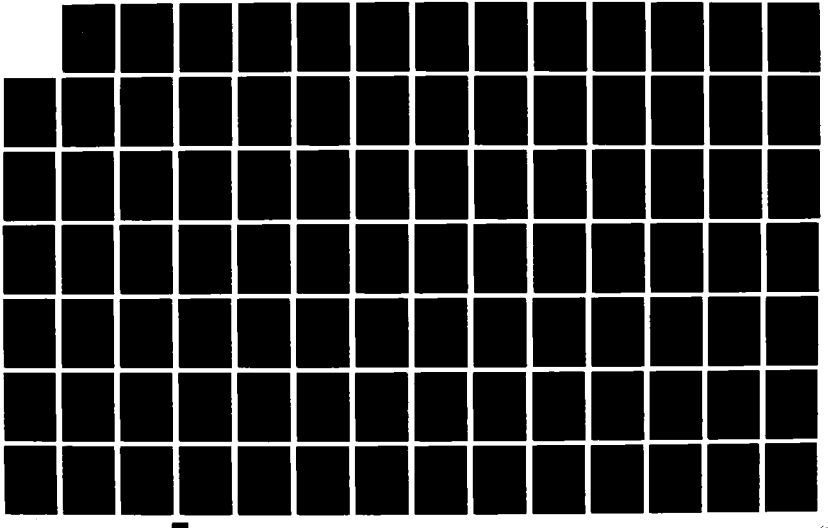
1. Introduction

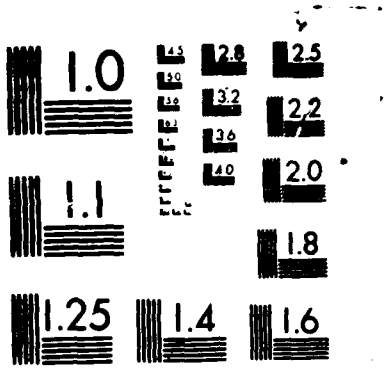
In the last three chapters, we have explored the integration of language analysis with the inferential memory processing required in order to understand. The point of such an integrated approach is to facilitate the early use of contextual information in the rational resolution of problems in language analysis, thus avoiding the indiscriminate use of backtracking that inevitably results from making such choices arbitrarily. But what about the inference and memory processes themselves? What implications does an integrated approach to language understanding have for such processes?

It is now widely accepted that understanding an input is fundamentally an issue of *explaining* that input in relation to its context and in terms of the understander's prior knowledge. This theory is often called *explanation-based understanding* (see, e.g., Schank and Abelson, 1977; Wilensky, 1978). Thus, for example, understanding a narrative entails forming an explanation that ties together the actions described in that narrative into a coherent structure (see, e.g., Charniak, 1972; McDermott, 1974; Schank, 1975; Schank & Abelson, 1977; Charniak, 1978; Cullingford, 1978; Wilensky, 1978; Dyer, 1983; Schank, 1982). In particular, understanding an agent's actions entails explaining why those actions might serve goals which could plausibly be ascribed to that agent. Understanding conversational input may additionally entail explaining why the speaker said what he said, or even why he said it in the way that he did (see, e.g., Schank, 1978; Cohen, 1978; Perrault, 1980).

Broadly speaking, an appropriate explanation of [illegible] consists of a chain of plausible inferences such that [illegible] from" certain other aspects of the context in which [illegible] the world. At each point in such a chain of [illegible] inferences that might be drawn next. [illegible]

AD-A183 553 INTEGRATED PROCESSING IN PLANNING AND UNDERSTANDING(U) 2/3
YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE
L BIRNBAUM DEC 86 YALEU/CSD/RR-489 N00014-75-C-1111
UNCLASSIFIED F/G 12/9 NL





explanation; the rest will not. Thus, explanation-based understanding, as much as language analysis or any other cognitive process, is confronted with the need to make decisions about which lines of inquiry are likely to prove fruitful, and should therefore be pursued. In other words, the basic problem which motivates an integrated approach to any cognitive process arises in explanation-based understanding as well.

In order to address this problem, an integrated approach to understanding must attempt to take as much relevant contextual information as possible into account, as early as possible, in the interpretation of an input. The most relevant possible contextual information is the set of contextual features -- the *hypothesis* -- in terms of which the input will ultimately be explained. Thus, in an integrated approach to understanding, the understander's hypotheses about the situation to which the input pertains must be brought to bear in deciding which inferences are likely to result in constructing an appropriate explanation, and should therefore be drawn from the input. Context must play a role in determining not only how an input should be interpreted, but also in how that interpretation is arrived at. In other words, an integrated theory of explanation-based understanding must be a relatively *top-down* theory.

The alternative, of course, is simply to search for an explanation by means of a purely bottom-up inference process, in which the choice of inferences to be drawn in connecting the input to an explanatory hypothesis is not sensitive to the presence of that hypothesis, at least not until the explanation has been found, at which point inference can cease. Thus, given the plethora of possible inferences which can be drawn from an input, or from a previous inference, such a model must choose which inferences to draw arbitrarily. As usual, two possible strategies are available to deal with the problems raised by the need to make such arbitrary choices, and the fact that most will be irrelevant or even misleading.

The first strategy is simply to accept that undirected inference will be performed, and to set about using a control structure that methodically searches the space of possible inferences, either in a breadth-first fashion (e.g., Rieger, 1975) or else depth-first with backtracking (as in, e.g., much work on plan understanding). In some sense, this model is a straw-man -- rejected even by many who have been forced to resort to it -- since it is widely acknowledged that, in an understanding system endowed with a realistic number of implication rules, such an approach is impractical, at least on a serial computer. Given a large branching factor, i.e., a large number of inferences that might be drawn from a given input, if the number of inferences in the explanation connecting an input with an hypothesis is even moderately long, then the space of inferences to be searched in constructing an explanation is extremely large. Moreover, on the methodological viewpoint taken here, the use of arbitrary choice and the resulting need for

brute-force search through the space of possible inferences is simply not a very interesting answer.

Because the problems of undirected inference are so widely understood, the more common approach to dealing with problematic decisions about which inferences to pursue is to avoid the problem by implementing exactly and only the inference rules necessary to explain the inputs that one knows, ahead of time, will be received -- in other words, the wishful control structure fallacy. In this way, one can implement a program that explains natural language inputs in a bottom-up fashion, while maintaining the illusion that undirected search has been avoided. Such a program *seems* to be a process model, but it is not. If the input is changed slightly so that another rule is necessary, the model would simply fail to understand. Adding all of the rules necessary to deal with small perturbations in the input would reveal that the model provides no theory of how to decide which rule to apply. Again, I must reiterate that I don't mean to imply that such models are entirely without merit: They often include interesting content theories of the knowledge that is necessary in order to understand a given class of inputs. But, construed as process models, they simply fail to address the issues that a *bona fide* process model must address.

At root, the wishful control structure fallacy stems from the limits of introspection. Since we ourselves already know the appropriate explanation for some input when we try to determine how our model should draw the necessary inferences to construct that explanation, it is relatively easy to come to see which inferences are required, and much harder to call to mind the inferences which are *not* relevant in this case. Indeed, this very fact provides phenomenological evidence for the claim that all inferences are not, and should not be, equally accessible to an understander in all contexts. However, it is one thing to attempt to explain how a functional model might achieve this result in general, and quite another to merely mimic it in a given example, by simply failing to include the rules that are not relevant to that example. Explanation-based understanding entails searching, hopefully in a directed fashion, the large space of inferences that could, in principle, be drawn from any input. A model which fails to come to grips with this fundamental fact is not, whatever other merits it may have, a serious process model of understanding.

In sum, short of either ignoring the problem or simply living with undirected inference in some way, it is clear that a theory of the inferential processes involved in explanation-based understanding must be a theory in which context can play a role in controlling inference. Indeed, it is well known that determining the relationship between a prior hypothesis and an input -- that is, constructing an explanation given both "endpoints" -- is computationally easier

than constructing the same explanation from the input alone. This can be illustrated, in its simplest form, by considering the obvious advantages of a *bi-directional* model of explanatory inference, in which inferences are drawn from both inputs and hypotheses until they "meet in the middle," resulting in a complete explanation.

There is much to recommend even such a straightforward model of contextual influence in explanatory inference over purely bottom-up inference from inputs alone, as a few simple calculations make clear. Given an average branching factor of **b** (representing the number of inferences that can be drawn, on average, from an input), and an average explanation length of **d** (representing the number of inferences that make up an explanation on average), then simply inferring bottom-up from an input would, on average, require on the order of b^d inferences to construct an explanation. Assuming that one had chosen the right hypothesis, on the other hand, bi-directional search would require about $2b^{d/2}$ inferences to construct the same explanation. Thus, the number of inferences required to construct explanations is reduced by a factor of two in the exponent -- which is a rather significant reduction. For example, if **b**=5 and **d**=4, then bi-directional inference would require a bit over 50 inferences to construct an explanation, whereas drawing inferences from the input alone would require over 625 inferences. If **b**=10 and **d**=4, then bi-directional inference would require somewhat more than 200 inferences, whereas inferring from the input alone would require well over 10,000 inferences.

In fact, the improved efficiency of bi-directional inference is significant enough that the extra cost of considering several hypotheses is usually worthwhile. That is, if bi-directional inference takes as its starting point **n** different hypotheses which might explain a given input, then the cost of constructing an explanation is about $(n+1)b^{d/2}$. As long as $n<b^{d/2}-2$, then, such a bi-directional inference process will still be more efficient than simply drawing inferences from the input alone. For example, if **b**=5 and **d**=4, then it is worth considering bi-directional inference from up to 23 different hypotheses. If **b**=10 and **d**=4, it is worth considering up to 98 different hypotheses. In fact, the savings is even greater than these calculations indicate, since the inferences drawn from different hypotheses about a situation can be saved for use with subsequent inputs pertaining to that situation, and the cost of such inference therefore amortized over all inputs to which the hypotheses might be applied.

Thus, the increased efficiency of bi-directional inference may turn out to be a sufficient account of the utility of context in understanding. Still, from the viewpoint of an integrated approach it seems a weaker theory than might be hoped for. It is true that, on this account, the inferences drawn from an input are in some sense contingent on the understander's

hypotheses, since they are contingent on inferences drawn from those hypotheses. In particular, most of the inferences that might have been drawn from an input after the point at which bi-directional inference discovers an explanation will not in fact be drawn. To some extent, then, bi-directional inference could be characterized as a top-down model of explanation-based understanding. On the other hand, the model is still bottom-up in the sense that the understander's hypotheses play no role in deciding which inferences should be drawn from an input until an explanation has been discovered. That is, although the inferences drawn from an input are contextually determined *as a class*, context plays no role in determining, on an individual basis, whether or not inferences within that class are likely to be useful.

Several motivations underlie the search for a more thoroughly integrated model of explanation-based understanding, in which the choice of inferences to be drawn from an input is more specifically determined by context. First, there is the introspective evidence, mentioned earlier, that all inferences that might be drawn from an input do not seem equally available in all contexts, which cannot be accounted for by bi-directional inference alone. Second, there is the assumption of mental determinism, which leads to the methodological principle that arbitrary choice in processing should be considered only when there seems to be no alternative. Thus, every effort should be made to find a model of explanation-based understanding in which the choice of which inferences to draw from an input is as rationally determined as possible, by taking into account the understander's hypotheses about the situation to which that input pertains.

Third, there is the hypothesis, previously discussed in chapter one, that the depth of inference required for understanding, and, most of all, the branching factor, are far greater than they are generally believed to be, so that even a bi-directional model of inference is likely to be overwhelmed. Indeed, I believe that the branching factor in explanatory inference is more like 50 than 5, and even this may be too conservative. Perhaps even more important than the *actual* branching factor -- which reflects only the number of implication rules that can be *successfully* applied to an input -- is the number of implication rules that might *potentially* apply to that input. That is, if the condition on the left-hand side of an implication rule is a conjunction, and the input satisfies one conjunct of that condition, then the rule is potentially applicable to that input. Whether or not the rule can be successfully applied -- that is, whether or not the inference specified by the rule's right-hand side is drawn -- depends on whether or not the other conjuncts in the condition hold. However, determining whether or not the other conjuncts hold will itself often entail substantial inference. To some extent, this distinction between potential branching factor and actual branching factor is an artifact of the inference engine employed. In a resolution theorem prover, for example, all rules would apply, but the

output of those depending on other conjuncts would simply be slightly less complex implications themselves. In this case, however, the actual branching factor would be correspondingly greater: In a sense, this model just makes explicit how expensive, in terms of depth and breadth of inference, it actually is to attempt to apply a rule with conjunctive conditions. Thus, regardless of whether or not a rule can be successfully applied, simply *attempting* to apply it may be computationally quite expensive. This argues that the attempt should not be made unless there is reason to believe that the results will be useful. In other words, the choice of which inferences to attempt to draw from an input should be, in part, contextually determined by the understander's hypotheses.

In order for such a theory to be functionally viable, several conditions must be satisfied. First, it must be reasonably easy to generate or retrieve the hypotheses most likely to explain a given input. Second, the inferences that are drawn from an input or from an hypothesis must be directed in some way by the fact that the task is not merely to draw inferences, but to draw inferences which relate the hypothesis and the input. In other words, the inferences that are drawn must be chosen because they hold out the best hope of explaining the input in terms of the hypothesis. Third, the total cost of these two processes, generating hypotheses and determining, in a directed fashion, whether or not they actually explain the input, must be less than the cost of even bi-directional search. That is, the total number of inferences drawn must be less -- hopefully, far less. These conditions are the *sine qua non* of a highly integrated theory of understanding.

In this chapter, we will concentrate primarily on the second of the above three conditions. That is, I will generally take it for granted that some mechanism for generating or retrieving likely hypotheses exists, and focus on the issue of how the presence of such an hypothesis can be used to direct inference. This is a significant simplification of the problem, but from a functional viewpoint, the second question is logically prior: If ways cannot be found to make good use of prior hypotheses in the explanation process, then there is no point in developing methods to retrieve them. The fundamental question to be addressed, then, is how it is that an hypothesis can be used to establish the inferences that are likely to be relevant in explaining an input in terms of that hypothesis.

One requirement on any model which seeks to answer this question is already clear: The cheaper the cost of gathering the evidence needed to determine whether or not a given direction of inference is likely to be useful in constructing an explanation, the better. Although in principle it is possible to develop a highly directed and efficient inference process in which determining the potential utility of a given inference involves drawing all possible inferences

from *it*, or determining the potential utility of an implication rule involves actually applying that rule and evaluating the result, such models will tend to be more useful in tasks where the problems of undirected search arise primarily from the *depth* of inference that is necessary, rather than the branching factor. That is, if long chains of inference are necessary in order to solve some problem, then the evaluation and selection of candidate directions for inference by means of a one- or two-ply "look-ahead" of undirected inference in all candidate directions can, if successful, result in eliminating a relatively large percentage of irrelevant inferences. The shorter the inference chains required, however, the smaller the percentage gains, and the larger the branching factor, the larger the total number of wasted inferences. Again, even more important than the actual branching factor is the number of rules that might *potentially* apply at a given point, and the cost of attempting to apply them.

In general, the explanations needed in order to understand the role of a given proposition in an established linguistic context do not appear particularly lengthy: As I have already stated, it is my belief that the problem stems primarily from the large number of inferences that can be drawn from any proposition, i.e., the branching factor, or, more exactly, the number of rules that might potentially apply. In fact, one of the features that distinguishes inference in domains with which an understander (or problem-solver) has a great deal of experience from those in which he does is that experience results in the creation of "macro-operators" (Fikes, Hart, and Nilsson, 1972) which encapsulate previously derived solution methods. Saving and using prior solutions in this way is fundamentally a trade-off in which depth of inference is potentially reduced at the expense of an increased branching factor. Thus, methods for directing inference which, as a matter of course, entail actually drawing the inferences that are to be evaluated, seem unlikely to be of much help in controlling inference in domains with which an understander or problem-solver is extremely familiar. In addition, such models do not seem capable of accounting for the introspective evidence that all inferences do not seem equally available in all contexts. These arguments suggest that, although the mechanisms involved in deciding which inferences to draw may themselves involve inference in weighing the evidence for or against the potential utility of some line of reasoning, the evidence itself must ultimately derive from some sort of non-inferential processing (see Charniak, 1983).

2. Script/frame theory

The sole current theory of understanding which meets the above criteria -- that is, in which the understander's hypotheses about the situation to which an input pertains are employed to determine specifically which inferences will be useful in explaining that input -- is *script/frame* theory (Minsky, 1975; Schank and Abelson, 1975 and 1977; Charniak, 1978;

Cullingford, 1978). The basic idea behind this theory, as a theory of processing, is as follows: In a given context of a sufficiently stereotypical sort, one can specify ahead of time the inferences which are most likely to be useful in explaining inputs in that context. That is, one can specify *expectations* about what sorts of inputs will be seen, and how to deal with them. In answer to the fundamental question posed above -- how can an hypothesis be used to determine the inferences that are likely to be useful in explaining an input -- script/frame theory postulates a very simple answer: associate the relevant inference rules *directly* with the hypothesis. If the script is *active* -- which is to say, if the state of being in the context associated with that script is believed to be the best hypothesis for the perceived inputs -- then only those inferences which are specified in the script should be applied to those inputs.

In a sense, the script/frame idea is a way of taking the wishful control structure fallacy and turning it into a theory. Instead of mistakenly assuming that the limited number of inference rules needed for a given example are the only ones that will ever be needed, it asserts that *if* the example falls into a natural class in which those are usually the only necessary inference rules, and *if* there is a simple way to detect whether the situation in which the understander finds itself falls into that class, *then* it is sensible to apply only those rules. To put this another way, script/frame theory asserts that a relatively simple context switching mechanism should be used to determine the set of inference rules that will actually be applied to inputs -- a set which is highly restricted in comparison to the number that might, bottom-up, be applied. In a sense, a script is to inference rules what its associated scriptal lexicon is to word meanings.

Script/frame theory is clearly a highly integrated theory of inference, in that the hypothesis which the understander brings to bear on an input -- namely, the script itself -- by definition delimits the set of relevant inference rules. But there is a second aspect to script/frame theory which is implicitly crucial to its successful application, and that is the extreme *specificity* of the expectations generated by a script. Each expectation generated by a script is, in effect, the left-hand side of an explanatory inference rule of the form "If you see this input, then interpret it as part of this script in this way." However, the conditions under which a given rule will apply -- its left-hand side -- are extremely specific and concrete. In script/frame theory, the explanation of an input can be accomplished with a single inference, if it can be accomplished at all. Expectations must be tailored to the exact form in which inputs are likely to be received if the script is to offer any guidance as to their explanation.

The need for this specificity follows from the simple method by which script/frame theory controls the inference rules that will be applied in a given context -- namely, by *directly*

specifying the relevant rules. Since these are the only rules that will be applied to inputs, then if the script is to be able to explain an input at all, one of these rules must be applicable to that input directly, by means of a simple, non-inferential, pattern-matching process. Otherwise, the script can offer no guidance as to how the input is to be explained. Therefore, the inference rules specified by a script must be tailored to the exact form in which inputs can be expected to be received.

In what form can inputs be expected? In general, linguistic and perceptual inputs are specific and concrete rather than general and abstract. Restaurant stories, for example, generally refer to the diner paying the check, not to the party of the second part fulfilling his contractual obligations, even though abstractly that is what is going on. Similarly, we see John and Mary hugging each other, not two old friends reconciling after a bitter argument, even though abstractly that may be what is going on. Scriptal expectations must be represented in the same concrete terms as inputs. If there is a mis-match -- if the expectations are too abstract in comparison with the inputs -- then determining which expectation applies to a given input, and how, will itself entail inference, and script/frame theory offers no guide as to how such inference can be accomplished in a directed fashion. Thus, the specificity of scriptal expectations is crucial to a script's ability to direct inference.

On the other hand, the very specificity of scriptal expectations that enables script/frame theory to focus the inference process necessarily limits its application, at least as a theory of *processing*, to highly stereotypical situations with limited variability (see, e g., Schank and Abelson, 1977, and Wilensky, 1978, for further discussion of this point). Highly variable situations, and especially novelty, necessarily require the application of more abstract causal knowledge in order to be understood, since the specifics cannot be predicted in advance. In such cases, script/frame theory is not, in general, sufficient, since the applicability of abstractly formulated explanatory inference rules to concrete inputs is not immediate and direct. Carried to an extreme, and applied to problems for which it is inappropriate, script frame theory degenerates into the wishful control structure fallacy, with an added twist: The inference rules employed are typically a "compilation" of the rules that should actually come into play in understanding the example, collapsing a chain of inferences into one rule that can be applied simply and immediately to the input, thereby maintaining the illusion of inference control. In other words, the misapplication of script/frame theory not only undermines the processing claims made by a model, but also distorts the way in which the model represents the content of the knowledge that it employs.

Despite these drawbacks, and the potential for misuse, there is a great deal of merit in

script/frame theory, if only because it is the only example of a theory of understanding in which the understander's hypotheses about a situation can determine the inferences which should be applied to inputs in that situation. It is, therefore, the only theory of explanation-based understanding that does not entail the use of arbitrary choice in inference. However, because script/frame theory depends on the use of highly specific expectations, it can only be used in understanding highly stereotypical situations with little variability, and no novelty. Understanding more complex situations requires the use of more abstract causal knowledge, as we will see more clearly in the next section. The challenge, then, is to develop a highly integrated theory of explanation-based understanding which is more flexible than script/frame theory.

3. Thematic knowledge in story understanding

Consider the following story:

> Smith and Wesson were competing for a government contract. Both of them needed the business, but Wesson was particularly strapped at the time. He competed fair and square, but Smith turned out to be extremely dishonest: He *bribed several government officials and won the contract for his company.* Wesson went bankrupt.
>
> Years later, Smith decided to run for Congress. So Wesson gave a lot of money to the Democratic party in his district.

This text exemplifies the need to use abstract thematic knowledge in story understanding. A reader would be able to understand the above story only if he understood that Wesson was attempting to get revenge on Smith by helping his opponent. A theory of how such stories might be understood must confront three questions: First, what is the content of the concept of revenge and how can it be represented? Second, how might a reader come to see that this concept is relevant to the text? And third, how does the reader come to recognize that Wesson's monetary contribution is in fact an act of revenge?

The above story has some special features that are particularly important in light of the last two questions. First, unlike what is typically the case with texts that can be understood using scripts or other highly specific memory structures, there is no word or phrase in the input that directly indicates the relevance of the revenge structure. For example, most texts that involve the restaurant script mention the word "restaurant," or perhaps a phrase such as "go out

eat," or "go out for Chinese," or something of the sort. In contrast, we cannot assume that some word or phrase in an input text will cue us in directly to the likely presence of an abstract thematic structure that we need in order to understand that text. Sometimes, of course, this does happen. That is, a variant of the above story might directly mention the word "revenge" in describing Wesson's campaign contribution. This would simplify our task, but the above story is perfectly comprehensible even without such a simplification. So we must assume that it is possible to access such structures even if they are not directly mentioned in the text.

The second problematic aspect of the above story is that giving money to a political party is a rather novel means of gaining revenge. Recognizing that such an action does in fact constitute an act of revenge is, therefore, not a simple inference. The difficulty posed by such novelty can best be understood in comparison with a more stereotypical variant of the above story in which the second paragraph, instead of describing Smith's run for Congress and Wesson's contribution, simply read "Wesson shot Smith to death." This simpler variant could, in principle at least, be understood using something like the following reasoning:

> Smith did something that drove Wesson out of business [given in the text]: That's usually a bad thing to do to somebody [inference]. Then Wesson shot Smith to death [given in the text]: *That's usually a bad thing to do to somebody too* [inference]. I know that I can explain some person X doing something bad to some other person Y if I know that X is angry at Y [causal knowledge]. So maybe Wesson is angry at Smith [inference]. I also know that I can explain some person X being angry at some other person Y if I know that Y did something bad to X [causal knowledge]. So maybe Smith did something bad to Wesson [inference]. Why yes, he did: He drove Wesson bankrupt.

The point here, of course, is that the reasoning described above is entirely bottom-up. That is, the inferences involved in explaining why Wesson shot Smith are not determined by any prior hypothesis based on Wesson's likely goals under the circumstances. Of course, such bottom-up processing is not very efficient: *The above description is a bit misleading in this* regard, because none of the irrelevant inferences that such a process would draw have been recorded. Since the explanation for Wesson's action is 4 inferences long, given an average branching factor of only 5, even bi-directional inference would produce more than 50 inf[illegible]nces in the course of its derivation. Literally following the reasoning described above, which is not really bi-directional -- 3 out of the 4 inferences in the explanation are generated from Wesson's shooting of Smith -- would require drawing about 130 inferences.

But the limitations of a bottom-up approach become even more apparent when we consider our original story. The crucial step in the above explanation of our variant story -- without which an explanation cannot be inferred -- is the realization that Wesson's act of shooting Smith to death is probably a bad thing to do to him. In the case of shooting someone to death, it is arguable that such an inference could be drawn bottom-up. But inferring that Wesson's monetary support for a political party will be bad for Smith is not even *possible* bottom-up, and even inferring simply that it will be bad for *somebody* requires an additional chain of several inferences, the construction of which is likely to swamp an undirected inference process. Indeed, it is clearly because we *already assume* that Wesson will try to gain revenge on Smith that we construe his contribution in this way, not the other way around. Thus, an adequate theory of the use of thematic knowledge in explanation-based understanding must be a *top-down* theory of understanding.

4. Representing thematic knowledge

Before proceeding with an attempt to answer the question of how thematic knowledge can be used in understanding, let's go back to the first question posed in the last section: How can we represent the structure and content of thematic knowledge? What are thematic structures about?

Thematic structures are structures concerned, not with particular goals and plans that arise in a situation, but more abstractly with *relations* among, and the types of, the goals and plans that arise in that situation (Schank, 1982). For example, a first pass at representing the concept of revenge would look something like this:

X does something bad to Y

CAUSES

Y is angry at X

CAUSES

Y wants to do something bad to X

CAUSES

Y does something bad to X

Figure 5-1: First pass at the abstract representation of "revenge"

The main point here is that the vocabulary involved in representing a thematic concept like

"revenge" is quite abstract -- it does not refer to particular actions or goals, but to classes of actions and goals and causal relations among them. The goal relational nature of the necessary representational vocabulary can be made even clearer by asking what it means to "do something bad" to somebody: It must mean something like to "block a goal of" that person's. So a more explicit representation of the concept of revenge would look something like this:

X blocks a goal of Y's

INITIATES

Y is angry at X

INITIATES

Y wants to block a goal of X's

REASON-FOR

Y blocks a goal of X's

Figure 5-2: Second pass at the abstract representation of "revenge"

Above, we have also made explicit the particular causal relations involved, using the causal taxonomy proposed in Schank (1975). This representation of "revenge" makes it clear that the vocabulary involved in the thematic level is one of causal relationships among goals and plans -- e.g., "blocking" various goals, and the causal relationships among various goals to block -- rather than specific goals and plans themselves.

Of course, we might wonder whether we actually need a specific revenge structure like the above in addition to the individual causal rules that seem to make it up. After all, we will in any case need such rules as "If X does something to block a goal of Y's, then Y will be angry at X," since we need to be able to understand and explain anger separately from revenge. How then can we justify the existence in memory of a revenge structure? The simplest answer would rest on an appeal to the increased efficiency that might result from "caching." That is, we freeze and save this particular cluster of inferences simply because it occurs frequently enough to warrant doing so (cf. our earlier discussion of "macro-operators").

There is a another argument we can make, however. The utility of large semantic structures stems not only from any functional role they may serve in inferential memory, but also because such structures generally carry with them extra information, above and beyond the causal "backbone" of elements from which they are composed, expressing the additional relations that can be expected to hold among those elements. To argue for the utility of an

explicit "revenge" structure in this way, we must answer the following question: What do we know about revenge above and beyond what is expressed in the smaller causal rules that relate goal blockage to anger and anger to blocking?

For one thing, we know that a person is much more likely to do something bad to someone else if they are angry at that person for doing something bad to them, than if they are angry for some other reason. That is, if Wesson were angry at Smith because Smith dealt dishonestly with a third party, Wesson would probably be less inclined to do something bad to Smith than if Smith had actually dealt dishonestly with Wesson himself. Of course, if Wesson has some close personal relationship with the third party, then he might very well do something bad to Smith -- this is a situation similar to revenge, expressed in English by the word "avenge."

Second, we also know that if Y wants to get revenge on X then the goal of X's that he will try to frustrate will be of roughly equal importance to the goal of his own that X managed to frustrate. For example, suppose that in our story, after Wesson realizes he is bankrupt, he sees Smith at lunch and starts screaming at him, and that as a result Smith's lunch is completely disrupted and he decides to leave the restaurant. Although Wesson did scream at Smith because he was angry, and although he did frustrate Smith's goal of eating lunch in that restaurant at that time, he probably didn't think of it as an act of revenge, nor did Smith, nor should we. In particular, Smith should still be on his guard with respect to Wesson, and we should still expect to see Wesson trying to gain revenge on Smith. The reason is simply that ruining someone's lunch is not likely to frustrate a goal that is nearly as important as going bankrupt does.

To sum up, we can see that an abstract thematic concept like revenge is composed of rules representing various causal relationships -- such as "blocking" -- between classes of goals and plans, and additionally imposes constraints on the particular goals and plans that are involved in any instance. The important point about such thematic structures from the point of view of explanation-based understanding is that the causal knowledge that they represent is extremely abstract. Thus, although such structures might be considered "frame-like" in their size and scope, they cannot realistically be employed in an integrated understanding system based on script/frame theory, construed as a theory of processing. The expectations that they generate are, for the most part, too abstract to be applied directly to inputs.

5. Prior work on thematic structures

The arguments I presented in the last section for the semantic utility of explicitly representing large, abstract thematic structures in memory are somewhat novel; but the idea itself is not. Abelson (1973) first proposed explicitly representing such structures in the course of investigating the representation of goal relationships in a belief system. He noted the existence of large thematic patterns such as "betrayal," some of which were so complex and yet so familiar that there are specific idiomatic phrases to express them in English. In addition to the concept of betrayal, for example, he proposed to represent such concepts as "the end of the honeymoon" -- which is to say, the stereotypic thematic pattern in which two agents decide to work together for common goals, and originally feel quite good about the relationship, but then eventually realize that they have conflicting goals as well, which ultimately come to the fore and thus change the feelings that the agents have towards one another. To take yet another of Abelson's examples, he proposed a representation for a thematic pattern which seemed to capture something of the notion of "the worm turns," i.e., the thematic sequence in which an agent who is dominated by -- or, in the vocabulary of goals, whose goals are constantly frustrated by or subordinated to those of -- another, more powerful agent, himself becomes, by some turn of events, the dominant agent.

Abelson's main point in this investigation was simply representational. That is, he simply wished to point out that patterns of goal relationships existed at this level of abstraction and complexity, and would need to be explicitly represented in a belief system which could reason about plans and goals, especially in the context of political questions. Additionally, he proposed some specific representations for these concepts in terms of a vocabulary of plan and goal relationships. Although quite advanced for the time, the specifics of the proposals were not particularly compelling because Abelson did not argue for their functional adequacy in any particular task such as planning or text understanding.

Lehnert's (1981) theory of *plot units* involves thematic concepts exactly like revenge -- in fact, revenge is one of the plot units she describes. Other structures of this type which she investigated include situations in which an agent is beset by the failure of an important goal, only to find that the situation presents an opportunity to satisfy some other important goal. Like Abelson, she noticed that such patterns often had a conventional rendering in language, in this case as "cloud with a silver lining." Yet another thematic pattern she described is the situation in which a planner pursues one goal only to find in addition that another goal can be achieved along the way as well (which she termed "killing two birds with one stone").

Lehnert was specifically concerned with the use of these abstract thematic structures in representing the meaning of texts. However, the use to which she put such structures was not in actually constructing the explanations which were required in order to understand, but rather in summarizing stories once they were already understood. In particular, the model which she proposed made use of the more atomic causal rules such as "If X does something bad to Y, then Y will be angry at X" in order to construct explanations in text understanding. Only after this was accomplished would the plot units which were constructed out of sequences of such atomic rules be recognized and then used in summarization. In other words, Lehnert proposed an entirely bottom-up theory of the recognition of abstract thematic structures, in which they played no direct role in story understanding.

Schank (1982) proposed a class of memory structures, *thematic organization points* (or TOPs), to express exactly the sort of abstract thematic structure relevant here. He proposed that the abstract causal sequence of goal and plan types that underly stories, jokes, fables, and so on, need to be explicitly represented in order to account for *cross-contextual* reminding -- that is, people's ability to be reminded of previous experiences or stories when presented with input situations that have the same underlying thematic structure, but are superficially quite different. For example, he investigated the thematic structure underlying such stories as "Romeo and Juliet," the apprehension of which enables us to recognize its similarity to another story such as "West Side Story." Other examples he analyzed included being reminded of such anecdotes as the story about the drunk searching under a lamppost for his keys even though he lost them elsewhere, "because it's where the light is."

In addition to a theory of how to represent abstract thematic structures, Schank investigated their role in understanding, learning, and planning. In particular, he wished to understand their role in memory organization, in order to account for the ability to retrieve cross-contextual remindings using the sorts of abstract features that make them up, and the use of such remindings in learning and planning. In order to play a role in reminding, such structures would clearly have to play a role in understanding. Thus, Schank argued that the representation of such abstract thematic properties of inputs was a necessary aspect of understanding. In addition, he pointed out the key role of expectation failures and the explanations for such failures in the formation and retrieval of thematic structures.

Dyer (1983) discusses an amalgam of Schank's proposals and Lehnert's. Like Schank, he proposes employing them in understanding rather than just in summarization. However, following Lehnert the model of understanding he proposes is entirely bottom-up. Thus, abstract thematic knowledge plays no role in determining how to infer the appropriate construal

of concrete input situations.

6. A closer look

Let's take a closer look at exactly what it is that makes our original story so problematic:

> Smith and Wesson were competing for a government contract. Both of them needed the business, but Wesson was particularly strapped at the time. He competed fair and square, but Smith turned out to be extremely dishonest: He bribed several government officials and won the contract for his company. Wesson went bankrupt.
>
> Years later, Smith decided to run for Congress. So Wesson gave a lot of money to the Democratic party in his district.

First, just in case it wasn't obvious to the reader at the begining of the chapter, Wesson's monetary contribution to the Democratic party only makes sense as an act of revenge if Smith is running for Congress as a Republican. This sort of inference -- finding the hypothesis which, if true, would allow us to explain what was going on in the story -- is called *abduction* (see, e.g., Pople, 1973; Charniak, in press). One of the problems posed by this example, then, is this: How is it that one generates an abductive hypothesis of this sort? It is perfectly tailored to bridge the gap between the hypothetical explanation -- revenge -- and the action to be explained -- Wesson's contribution to the Democratic party. Furthermore, the abductive inference that gets drawn is the one that leads to the best explanation overall for the situation. As we discussed in the previous chapter with regard to the problem of lexical ambiguity, it is not enough to find *an* explanation for some input. There are many possible explanations for any action or any aspects of an action. In fact what seems necessary for understanding is the ability to infer the *best* explanation, in terms of such criteria as coherence, parsimony, specificity (accounting for the most details), and so on.

For example, suppose we knew that Wesson was a Democrat in our original story. Then one possible explanation for why he made the contribution would simply be that he wanted to help his party. But this explanation would not render the story coherent, because it would not connect Wesson's contribution, Smith's candidacy, and their previous relationship into a unified explanatory structure. Nor would it explain why Wesson made the contribution in any particular district. In other words, the explanation of Wesson's contribution in terms of

revenge explains *more* of the story, both in breadth of coherence and depth of detail, than the explanation in terms of his political beliefs. That is what makes it a better explanation.

Before we can understand *how* the inferences that make up such an explanation get drawn, however, we must first determine exactly *what* those inferences are. The best way to answer such a question is to suspend our worries about the wishful control structure fallacy, and sketch out a "just so" story in which the inferences are drawn in a purely bottom-up fashion, ignoring the problem of how those particular inferences should be chosen. There is nothing wrong with this, so long as we don't confuse the resulting "protocol" with a *bona fide* process model, for example by directly implementing exactly and only these inferences in a program. In any case, the necessary inferences seem to be the following:

(0) Wesson gave money to the Democratic party in some district.
(1) The Democratic party may use this money to help elect Democrats to public office in that district.
(2) Wesson's contribution may help Democratic candidates in that district to win.
(3) The contribution may tend to cause Republican candidates in that district to lose.

So far, we have sketched out the necessary inferences as if they were entirely bottom-up. At this point in the explanation, however, we need the abductive inference that Smith is a Republican. If we simply inserted this inference into the flow at this point, we could continue constructing the explanation in a bottom-up fashion, as follows: "Smith is the Republican candidate, therefore, Wesson's contribution may tend to cause Smith to lose. Smith probably wants to win, therefore, Wesson's contribution may tend to block Smith's goal of winning." At this point, then, it would be possible to recognize an instance of revenge. It should be clear, however, that we cannot expect the abductive inference to just "pop up" in this way. In the strictest sense, then, it is not possible to explain this input in terms of revenge in a purely bottom-up fashion -- the abductive inference must take the explanatory context into account.

However, it is possible to pursue a bottom-up explanation process in which the abductive inference is drawn only at the very end, when the explanation is complete. Such a process will draw essentially the same inferences, except without any reference to Smith, as follows: "(4) The Republican candidate probably wants to win, therefore, Wesson's contribution may tend to block the Republican candidate's goal of winning." At this point, in other words, even though we don't know whose goal it is that Wesson's contribution may tend to frustrate, we can see that it will tend to frustrate *someone's* goal to win, and that that

someone is a Republican. All that is required, therefore, is to allow this inference to match the hypothesis, derived from the revenge structure, that Wesson will try to block a goal of Smith's, even though Smith is not mentioned in the inference. Such a matching process would entail drawing, either implicitly or explicitly, the abductive inference that Smith is the Republican candidate (see Charniak, in press, for a discussion of abductive unification).

In principle, therefore, it is possible to nearly construct an explanation in a purely bottom-up fashion, postponing all abductive -- and hence context-dependent -- inference until the very end, when all the inferences are assembled into an explanation. The problem is, of course, that at every step along the way in the above "just so" story, there are an enormous number of inferences that could have been drawn, but that would not be relevant to the explanation in this case. We cannot simply make it impossible to draw those inferences, because they might be necessary in some *other* case. It is instructive to try and write down some of the other inferences that might actually be drawn from Wesson's monetary contribution to the Democratic party:

Wesson is a Democrat.
Wesson is a liberal.
Wesson wants the Democratic candidate to win.
Wesson wants the Republican candidate to lose.
Wesson wants the Democrat candidate to like him.
Wesson wants the Democratic candidate to help him win government contracts.
Wesson believes the Democratic candidate will vote as he would on certain issues.
Wesson believes the money will help the Democratic candidate.
Wesson believes the Democratic candidate needs money.
The Democratic candidate may use the money to pay campaign expenses.
The Democratic candidate may use the money illegally for his private purposes.
Wesson has less money.
If Wesson was thinking of doing something else with the money, he can't.
Wesson had enough money to make the contribution.

Even when it is clear just how many plausible inferences really can be drawn from a given input in conjunction with background knowledge, it is often difficult to imagine circumstances under which those inferences will actually be necessary in order to understand that input. It must be emphasized again, therefore, that *any* inference that can be drawn from an input may be crucial to the explanation of that input in some context. Any of the inferences listed above, and many more besides, could be rendered necessary for understanding by the

construction of a suitable context. For example, consider the following three stories:

> Wesson was worried about the situation in Central America. So he gave a lot of money to the Democratic party in his district.
>
> Wesson was worried that his children would turn out to be spoiled brats. He decided to give a lot of money to charity.
>
> Elderly Wesson gave a large sum of money to the Democratic party. His heirs were very upset.

An explanation of Wesson's contribution in the first story must employ the fact that Democrats tend to be liberals to infer that Wesson is probably a liberal, and to then infer that his concern with Central America probably stems from his fear that the Reagan administration's policies might lead to U.S. military involvement in that region. In order to understand the second story, the reader must infer that Wesson is giving his money to charity so that he will have less of it, which means, in turn, that his children will inherit less, and will thereby be forced to become more self-reliant. Notice also that in this case, the inference that Wesson's gift will help those charities to achieve their goals is not particularly salient. The third story is a variation on the second, in which, although Wesson's intent is not necessarily to have less money, that result of his actions must be inferred in order to understand his heirs' reaction.

Let's consider another variant of our original story, in which the action that Wesson takes is to call the New York *Times*. The most likely explanation in this case is that Wesson is calling a reporter at the *Times* to tell him about Smith's disreputable -- in fact, illegal -- business conduct. The purpose, presumably, is to try and block Smith's election by publicizing his dishonesty, or even more, attempting to get him indicted for his illegal business activities. In fact, the story need not even involve a Congressional race by Smith. If, after describing how Smith bribed government officials and drove Wesson bankrupt, the story simply concluded with Wesson calling the New York *Times*, we could still understand this as part of an attempt to retaliate in some way against either Smith or the government officials who were bribed by him. Constructing that explanation entails the use of very specific knowledge of the use of the press to publicize wrongdoing or, in a less moral vein, simply to hurt someone's reputation. That is, publicizing misconduct (or alleged misconduct) must be a known, stereotypical plan in order for this variant story to be understood -- stereotypical in the way that shooting someone to get revenge is, and as making a political contribution, in contrast, is not.

The point here is that even when such stereotypical knowledge is available, it is not necessarily relevant. A bottom-up approach to explanation would, presumably, always infer that Wesson might want some information publicized. And, indeed, if a person contacts a professional or an organization, there is a default implication that he may have some business involving the primary function of that professional or organization. Nevertheless, on a highly integrated view of understanding, even such default inferences need not *necessarily* be drawn. For example, consider the following text: "Joe was a salesman for the North American Paper company, in charge of selling newsprint in the Northeast. He called the New York *Times*." Here, I would argue, we do not infer that Joe is calling to publicize some information. Similarly, in a context in which two people were arguing over some fact, and one of them called the *Times*, we would presumably immediately infer that he wished to talk to an authority there to settle the dispute, and not that he wished to publicize something.

7. Extending the range of integrated understanding

The central issue to be addressed by an integrated theory of explanation-based understanding is the question of how the hypotheses that might explain an input can help in the explanation process itself. This question can be usefully reformulated as follows: What *information is associated with an hypothesis that might be useful in explaining an input*? That is, if the query "How could input X be explained by hypothesis Y?" is easier to answer -- that is, entails far less inference -- than the query "What is the explanation for input X," then it must be because of some explanatory knowledge associated with the hypothesis Y. But in order for any of this knowledge to be at all useful, there must be some way to apply it in the explanation of the input.

Bi-directional search and script/frame theory, for example, are two proposals as to how the information associated with an hypothesis might be applied in explanation. In the former, the knowledge is used to draw inferences from the hypothesis itself. For example, if the hypothesis were an agent's goal, and the input were some action performed by that agent, then bi-directional search would proceed by inferring not only plans and goals that might entail the performance of that action, but also plans and sub-goals that might be used to carry out the hypothetical explanatory goal. In script/frame theory, the knowledge associated with the hypothesis is in the form of specific expectations that can be applied directly to inputs.

However, as we saw earlier in the chapter, this will only be possible if the expectations are tailored to the exact form in which the inputs are actually received. Except in highly restricted circumstances, this will not be the case. For example, if the hypothesis is that the

agent has the goal of gaining revenge, then, although there is a great deal of causal knowledge associated with this goal, that knowledge is highly abstract. As a result, the expectations generated by such an hypothesis cannot be applied directly to most inputs, because such inputs will generally be couched in a specific, rather than abstract, form. In other words, there is a mis-match between the representational vocabulary in terms of which inputs are expressed, and the representational vocabulary in terms of which the information associated with the hypothesis is expressed.

Thus, if the causal knowledge associated with an hypothesis is to be useful in explaining an input, then either that knowledge must be transformed into terms commensurate with the input, the input must be transformed into terms commensurate with the hypothesis, or else both must be transformed into commensurate terms. In bi-directional inference (or, for that matter, purely bottom-up processing), such a transformation is effected gradually, in the course of the inferential processing that actually constructs the explanation. We might ask, however, whether that transformation could be accomplished more directly, in order to facilitate the more immediate application of information associated with the hypothesis to the explanation of the input. In other words, we can conceive of explanation-based understanding as involving a *translation* process of sorts, by means of which input and hypothesis are represented in commensurate terms.

Because the explanatory knowledge associated with an hypothesis is likely to consist of quite a large number of causal rules, it will in general be more efficient to transform the input into terms commensurate with the hypothesis, rather than transforming all of the potentially relevant causal rules into terms commensurate with the input. Even so, of course, such a translation process will itself entail a certain amount of inferential processing. If the only way that it can be accomplished is by drawing inferences from either the input, the hypothesis, or both, in an essentially undirected fashion, then there is nothing to be gained by viewing the understanding process in this way: the resulting model would be bi-directional search. Thus, in order for this approach to succeed, it must be possible to accomplish the translation process itself in a fairly directed fashion. I will return to this point below. First, however, let's investigate how such a translation process might be useful in explanation-based understanding.

For example, consider our original story above involving Wesson's attempt to gain revenge on Smith. The input of interest is Wesson's monetary gift to the Democratic party. The hypothesis in terms of which this input should be appropriately explained is that Wesson has the goal of gaining revenge on Smith for Smith's prior role in driving him bankrupt. In particular, the expectation is that Wesson may try to block an important goal of Smith's. The

understander must now attempt to transform the input -- Wesson's gift -- into a representation commensurate with this hypothesis, that is, into an abstract characterization in terms of causal relations among goals. In this case, an appropriate characterization of the input in these terms is that giving money to an agent is a way of helping that agent achieve its goals.

Once the input has been translated into terms that are relevant to the hypothesis, the knowledge that is associated with the hypothesis can be brought to bear. In this case, the hypothesis is that Wesson's contribution is in service of a goal to block an important goal of Smith's. What do we know, in the abstract, about how to block another agent's goal? In fact, we possess quite a lot of such *counter-planning* knowledge (see Wilensky, 1978, and especially Carbonell, 1981, for discussions of counter-planning). We know, for example, that an agent's goal can be blocked by disabling a precondition for its achievement, by threatening another goal that is more important to the agent (so that he must divert resources from the pursuit of the first goal in order to protect the second), by "beating him to the punch," and so on.

Given this sort of abstract causal knowledge associated with the hypothesis, an understander can now attempt to apply it to the transformed representation of the input. In effect, this entails posing a query about the understander's counter-planning knowledge: Is there any way in which *helping* an agent achieve its goals could be used in order to *block* an agent's goal? The relevant rules, if they existed, could be retrieved using a variety of indexing techniques, including such stalwarts as discrimination trees or hash tables. In this case, in fact, there are several rules that could be retrieved as potentially relevant. One possibility is that the agent being helped has been recruited specifically to block the other agent's goal. Another rule, and this is the one that matters in this instance, is that helping an agent achieve its goals can be used to block another agent's goal if the agent being helped is in *competition* with the other agent for that goal ("the enemy of my enemy is my friend"). Attempting to apply this rule would, in turn, result in the following query: Is the agent being helped -- i.e., the Democratic party -- in competition with the agent whose goal is to be blocked -- i.e., Smith? The answer is yes, by the following reasoning: Smith is running for Congress, therefore he is in competition with another agent to win the election. Now, if the understander forms the abductive hypothesis that the agent with whom Smith is in competition is the Democratic party, then, since there are only two major political parties in the U.S., it follows that Smith is running as a Republican. Assuming these abductive inferences, then, Wesson's contribution can be seen to serve the goal of blocking Smith's election.

Let's look at some other examples, also involving a monetary gift to an organization, to

see how different inferences would be relevant in a different context. Consider the following story:

> Fred's accountant told him that the I.R.S. was going to slaughter him if he didn't act quickly. So he decided to give a lot of money to Planned Parenthood.

Here, the goal that is aroused by the first sentence is "reduce taxes." How could a monetary gift to Planned Parenthood reduce taxes? Presumably, indexed directly under the goal "reduce taxes" is the plan "Reduce taxable income by giving money to a non-profit organization." In other words, the fact that Fred's gift implies that Fred has less money is actually important to the explanation in this case. Compare this situation with the following:

> Fred was concerned with keeping abortion safe and legal. So he decided to give a lot of money to Planned Parenthood.

Here, the goal is to promote certain social goals. As in the story of Wesson's attempt to get revenge on Smith, the input must be transformed into a more abstract representation, namely, that Fred's contribution will help Planned Parenthood to achieve its goals. Now we can ask the following question: How could a contribution to a non-profit organization like Planned Parenthood promote certain social goals? The answer is, if Planned Parenthood promotes the same goals. Notice that in this case, no inference would be drawn about how the contribution to Planned Parenthood affected Fred's finances.

The approach sketched out above can be viewed as an extension of script/frame theory in several ways. One way to view it is as an attempt to increase the flexibility of script/frame theory by increasing the flexibility of the manner in which a script specifies the inferences that will be relevant. That is, on this view, not all of the relevant inferences need be *directly* specified, ahead of time, by the hypothesis (script) itself. Some, in particular the inferences involved in transforming the input into a form commensurate with the hypothesis, will be chosen on a more dynamic basis. Of course, as I have already pointed out, much then rides on the way in which the translation process is accomplished. If that should turn out to require undirected inference, then this approach reduces, at best, to a hybrid of script/frame theory and bi-directional search. We will return to this point shortly.

The other way to view this proposal as an extension of script/frame theory is to view the translation process as being instrumental to a more sophisticated form of pattern matching than is usually employed in applying scriptal expectations to inputs. On this view, the explanatory

inference rules directly associated with an hypothesis are applied to inputs as in script/frame theory, with the difference being that the pattern-matcher which applies them is capable of resolving any mis-match that might be due to the expression of input and explanatory rule in incommensurate terms. As before, however, we must be extremely careful about the means by which such an "intelligent" pattern-matcher will accomplish the translation. The simplest way to implement such a scheme, after all, would be to use a theorem-prover as the pattern-matcher. In this case, once again, the resulting process would be equivalent, at best, to bi-directional search.

There are, in addition, several reasons why it might be better to view the translation process as separate from the process of rule application *per se*. (An integrated approach is not, after all, fatally opposed to the idea that cognitive abilities are carried out by separate modules. It does insist, however, that such modules be *functionally* determined, and that they take into account as much information as seems necessary in order to perform their functions, rather than being defined by rather arbitrary *a priori* notions about what ought to constitute the source and type of information with which they should be concerned.) First, and simplest, each attempt to apply an explanatory inference rule to an input should not have to recapitulate the work of transforming the input into terms commensurate with the hypothesis. Second, the attempt to apply an explanatory rule to an input may involve the need to perform a great deal of other inferential work beyond what is needed to translate the input, for example, in forming abductive assumptions. To the extent that such processing takes place before it is determined whether the input can even be transformed into what is necessary in order to successfully apply the rule, it will be wasted if the result is negative.

Both of these problems, in turn, are related to the third, and perhaps the most important issue, which is how to determine which of the explanatory inference rules associated with the hypothesis are likely to be useful. There is a great deal of knowledge associated with most hypotheses, and knowing which of it is relevant is crucial to developing an efficient, directed inference process. In script/frame theory, this issue is not particularly pressing, since the process of rule application is not very expensive, and it is therefore quite feasible to attempt to apply all of the rules to every input. However, if the pattern matching process itself entails the use of inference, the cost of an unsuccessful attempt to apply a rule can be quite high. It is therefore important to develop computationally inexpensive methods for narrowing the set of rules which are likely to be applicable.

In fact, in script/frame theory -- where it is least needed -- this can be accomplished quite easily. Since the form in which inputs will arrive is already known ahead of time, one can

index the expectations in a script according to features of the inputs to which they are likely to apply. Then, using any of the standard retrieval techniques, e.g., discrimination trees or hash tables, one can retrieve only those expectations that are likely to apply to the input. However, if the explanatory hypothesis is more abstract, then by definition one does not know ahead of time the exact form of the inputs to which it might apply. Therefore, given an untransformed input, one cannot use any simple indexing techniques to retrieve candidate rules, since the input will not generally contain the features which might have been used to index those rules in the first place. If, however, the input has already been transformed into a representation commensurate with the hypothesis, then of course it *will* contain the sorts of features which could be used to index the rules ahead of time. Any one of a number of standard indexing techniques could then be used to cheaply retrieve the explanatory rules that are likely to apply to such a transformed input. Thus, for example, in the explanation process sketched out above, it was not necessary to attempt to apply *all* of the understander's counter-planning knowledge to the input (Wesson's monetary contribution to the Democrats), but only those rules that involved helping another agent achieve its goals.

Although the characterization of an input in terms commensurate with an hypothesis is a necessary prerequisite for applying the causal knowledge associated with the hypothesis to the explanation of that input, it is not a panacea. Even if this translation can be accomplished rather easily, the resulting representation may prove difficult to interpret. For example, consider a variation on our original story, in which Wesson's response to Smith's candidacy was to buy the main TV station in his district. This story is, I think it is fair to say, rather difficult to understand. However, one possible explanation of Wesson's action in terms of revenge would be that he intends to use his control of the TV station against Smith's candidacy by denying him favorable media coverage, while providing it to his opponent.

To be more specific, the counter-planning rule that is being used here is that one way to block an agent's goal -- in this case, Smith's goal of being elected -- is by blocking one or more of his sub-goals -- in this case, his goal of gaining favorable media coverage. However, there are several other counter-planning rules, of increasing specificity, involved in constructing a complete explanation of Wesson's action along these lines. The particular method by which Wesson is attempting to block Smith's subgoal is by blocking his access to a necessary resource -- the TV station, and in particular, its editorial staff -- for achieving that subgoal. And, finally, the particular method by which he is attempting to block Smith's access to the resource is by gaining control over the resource himself -- i.e., by buying the TV station.

Thus, the appropriate characterization of Wesson's purchase of the TV station -- in terms

of the representational vocabulary used to express counter-planning knowledge -- is as an attempt to gain control of a resource. The problem is that however easy it might be to derive this characterization, it provides very little guidance to the understander as to how to proceed in explaining it. That is, gaining control of a resource is such a ubiquitous activity in planning and counter-planning that Wesson's action could be a part of a plan to block Smith's goal in innumerable different ways. In sum, although translating the input into terms commensurate with the hypothesis may be a necessary step in explanation-based understanding, it does not ensure that an explanation can be easily constructed in all cases. Even when properly characterized, an input may be difficult to explain.

8. Transforming the hypothesis

In the last section, I sketched out an approach to integrated understanding in which the input is first transformed into terms commensurate with the hypothesis. In some cases, however, it may be more appropriate to do things the other way around, by transforming the causal knowledge associated with the hypothesis into terms commensurate with the input. This will generally be the case when such a transformation can be accomplished with little or no inference, for example, when a similar transformation has been previously performed, and the results have been stored in memory and can be retrieved on the basis of some *fairly simple* cues -- that is, when there is a *script* that might apply to the current situation.

Suppose, as before, that the understander has formed the hypothesis that Wesson may seek to gain revenge in attempting to explain the situation that is unfolding in the story. How might Smith's candidacy relate to this hypothesis? The understander already believes that Wesson may seek to thwart a goal of Smith's. But, as I pointed out earlier, not just any goal will do: The goal to be thwarted must be of roughly equal importance as the goal which was originally thwarted. Since Smith's goal of winning an election to Congress is of roughly equal importance as Wesson's goal of maintaining a successful business, the understander can transform the expectation that Wesson may seek to thwart an important goal of Smith's into the more specific expectation that Wesson may seek to thwart this particular goal of Smith's.

The formation of a more specific hypothesis of this sort may well be more useful than the general hypothesis from which it is derived. This will be particularly true if the understander already possesses specialized counter-planning knowledge of how to block someone's election to Congress -- that is, a script. This is not to say that that the understander necessarily *will* have such a script, or that it will necessarily cover the situation that arises. But if these conditions hold, then the inference process that would otherwise be necessary to

transform the input into terms commensurate with the hypothesis of revenge would be unnecessary. In effect, rather than translating the input into terms commensurate with the hypothesis, the understander would be translating the hypothesis into terms more likely to arise in the input.

In this case, for example, an understander might already know that one way to block a candidate's election is to give money to his opponent, or by promoting unfavorable publicity about the candidate. Thus, even though such specific expectations about Wesson's behavior would not be part of the understander's general counter-planning knowledge, it would be possible to retrieve them after determining that Smith's candidacy was a likely target for Wesson's revenge.

In principle, of course, it might even be possible to *invent* such specific counter-planning methods if they were not known ahead of time. That is, by applying general counter-planning knowledge to specific knowledge about how Smith would go about getting elected, an understander could determine how Wesson might attempt to block that goal. (This notion of applying abstract thematic structures to specific situations in order to derive instances of generic plans within a particular domain is currently being pursued by Collins, in preparation, in the domain of planning in competitive situations). However, in this case the task of the understander would be as difficult as the task of a planner attempting to create such methods for the first time. That is, if an appropriate script does not already exist, the attempt to create one top-down will require a great deal of inference, most of which will probably be wasted.

9. How to transform the input

So far, we have argued that the integrated use of an abstract thematic structure such as revenge in explanation-based understanding will often entail the need to somehow transform the representation of the input into a form which can be usefully related to such a structure. If this problem turns out to be as hard as the original inference problem itself -- that is, if such a transformation can only be accomplished by undirected inference -- then we will be no better off than we were when we started. The success of this approach depends, therefore, on finding ways to direct the inferences required in order to properly characterize the input. In other words, the process of reformulating an input in terms commensurate with an hypothesis must *itself* be an integrated process. The context provided by a potentially explanatory hypothesis must somehow be taken into account to help determine the appropriate representational vocabulary in terms of which to characterize the input, that is, the features of the input situation which are likely to be important in that context. Indeed, in view of the fact

that the number of different ways to construe any given input seems enormous, this is perhaps the most important function of such contextual knowledge.

In the simplest case, the specification of the representational vocabulary in terms of which an input should be represented could be accomplished *directly*, with a sort of scriptal "lexicon." That is, on this view an hypothesis would directly specify a set of features that could be used to determine, in a non-inferential fashion, a set of inference rules that, if applied to an input, would be likely to result in an appropriate abstract representation of that input. For example, let's consider trying to transform our example "Wesson gave a lot of money to the Democratic party" into "Wesson did something to help the Democratic party achieve its goals." In a sense, it's obvious that giving money to an agent will usually help that agent achieve its goals. There is, clearly, an implication rule to the effect that giving something to an agent is a way of helping that agent to achieve its goals, if the object given is a resource that could be useful in achieving those goals. But the question is, why do we choose to apply this particular rule and draw this particular inference, rather than others?

The answer, according to the simple model proposed above, is just this: Since the goal is to transform the input into the vocabulary specified by the hypothesis, the first step should be to check whether there are any rules that might do so directly. Such rules would have the property that the conditions on their left-hand side would be expressed in terms of the same representational vocabulary as the input, while the consequences on their right-hand side would be expressed in terms of the representational vocabulary specified by the hypothesis. Now, if the implication rules applicable to input concepts were indexed under those concepts in terms of the representational vocabulary used to express the consequences on their right-hand sides, then the relevant rules could be retrieved quite easily. The understander would simply use the representational vocabulary specified by the hypothesis as keys to retrieve the rules, if any existed, that were indexed under the input in those terms. Any rule that was retrieved would be potentially useful in transforming the input into a form commensurate with the hypothesis.

In this particular case, the keys provided by the revenge structure would include the representational vocabulary in terms of which counter-planning knowledge is expressed. That knowledge includes, for example, rules for blocking an agent's goal by disabling a precondition for its achievement, by threatening another goal that is more important to the agent, by "beating him to the punch," and by helping another agent with whom the first agent is in competition for the goal, among many others. The representational vocabulary involved includes, then, "disable," "precondition," "threaten," "goal," "help," "competition," and so on. Some of these representational elements, e.g., "precondition" and "goal," are so general that

their use as indices would probably result in the retrieval of many irrelevant rules, and this should probably be taken into account in the retrieval process. Using only the more specifically relevant items as keys to retrieve rules indexed under the input "give money" would result in the retrieval, perhaps among others, of the rule that giving something to an agent is a way of *helping* that agent achieve its goals, if the object given is a useful resource in the achievement of those goals. Applying this rule to the input would lead to the inference that Wesson's contribution was a way of helping the Democratic party achieve its goals, since money is an almost universally useful resource. The explanation of this more abstract characterization of the input would then proceed as described previously.

To take another example, consider the variant story in which Wesson buys the TV station in Smith's district. In this case, the relevant characterization of the input is that Wesson now has control of the TV station. The implication rule that generates this inference can be retrieved rather easily, since "gain control" is one of the representational vocabulary items in terms of which counter-planning knowledge is expressed. (Unfortunately, as I pointed out earlier, even though this characterization of Wesson's action is easy to derive, it is still quite difficult to explain.) In fact, it might be argued that this inference should always be drawn from such an input. However, consider the following story: "Fred had a lot of extra money lying around. He hated to see idle capital, so he decided to buy the main TV station in town." Here, it seems that the explanation for Fred's purchase in terms of his goal of making money need not entail drawing inferences about his control over the TV station.

Thus, as long as there are rules indexed directly under the input in terms of the contextually relevant features -- rules, in other words, which are likely to produce an appropriate characterization of the input -- then the direct specification of those features by the hypothesis, in conjunction with any of the standard indexing and retrieval techniques, can be employed to retrieve those rules. It is therefore possible to translate the input into terms commensurate with the hypothesis in an efficient and directed fashion, by using simple, non-inferential retrieval techniques to narrow the set of relevant implication rules. The importance of narrowing down the set of potentially useful rules is particularly great because, as I pointed out earlier in the chapter, the attempt simply to apply these rules often entails further inference. For example, the implication that giving something to an agent is a way of helping that agent achieve its goals depends on the further condition that the object given would be useful in achieving those goals. The attempt to verify such conjoined conditions may be computationally quite expensive, regardless of the outcome, which means that any non-inferential means of narrowing down the set of rules to be applied will result in large savings. And, as we saw previously, the appropriate characterization of the input enables a

similar use of non-inferential indexing techniques to narrow the set of potentially applicable explanatory rules once the translation has been accomplished.

These techniques can be applied in somewhat more interesting cases as well. Consider yet another variant of our original story, which simply ends as follows: "Wesson decided to become more involved in the Democratic party." Here, once again, the inference that must be drawn in order to understand this action in terms of revenge is that Wesson's newfound political activism is aimed towards helping the Democratic candidate for Congress, and, therefore, against Smith's candidacy. But as before, this characterization does not seem necessary in all contexts. Consider, for example, the following two stories:

> Fred was lonely and wanted to meet new people. He decided to become more involved in the Democratic party.

> Fred was ambitious and wanted power. He decided to become more involved in the Democratic party.

In both of the above stories, although it may be true that Fred will help Democratic candidates, this is obviously not germane to the explanation of his decision to become more politically active. Instead, in the first story, it is clear that the relevant characterization of the input is in terms of its social consequences: Political activism entails participating in group efforts, and involvement in group efforts offers a good opportunity to develop personal relationships with other members of the group. In the second story, it is clear that the relevant characterization of the input is in terms of its professional consequences: The way to gain a position of authority in a certain field is to become known and respected by other people of authority in that field, and one way to do that is to work with them.

The problem that we face, then, is the following: "Become involved" is a very vague locution, meaning, at best, something like "engage in actions causally relating to." In other words, the problem posed by this input is not that it is too *specific* with regard to the causal rules associated with the hypothesis, but that it is too *general* (cf. our discussion in earlier chapters of the problems posed by words with vague and general meanings). We might assume the existence of an implication rule asserting that if an agent engages in actions causally relating to another agent, then he might be engaging in actions to help achieve the goals of that agent. Assuming the existence of such a rule, after all, amounts to no more than assuming the existence in memory of a category of "actions which are causally related to some agent," and of "actions which help to achieve the goals of an agent" as a subtype of that category. The

problem is that it is hard to imagine a counter-planning method which *cannot* be characterized as a subtype of the category "actions which are causally related to some agent." In other words, the difficulty posed by this input is not so much that it is represented in terms which are incommensurate with the hypothesis, but rather that it is not very useful in narrowing down the set of relevant explanatory rules associated with that hypothesis.

However, in this case the input also contains more specific information. It specifies, in particular, that Wesson has decided to "become involved" in the Democratic party. Suppose now that we extend the methods described above, just a bit, so that the understander uses the keys supplied by the hypothesis to index under *all* of the concepts mentioned in an input, not just the main action. Applying these keys to "Democratic party" will not retrieve an implication rule *per se*. However, since one of the keys is "competition," it might retrieve a relevant *fact*, namely, the fact that the Democratic party is in competition with the Republican party. This would have two important consequences. First, it would enable the understander to characterize the input as meaning that Wesson has decided to become involved with a group, the Democrats, who are in competition with another group, the Republicans. Second, and most important, such a characterization in terms of the feature "competition" would enable the retrieval of the relevant counter-planning rule in this case -- that one way to block an agent's goal is by helping another agent who is in competition with him for that goal.

The use of standard indexing techniques to retrieve useful rules and facts as described above depends on such rules and facts being indexed directly under the input concepts in terms of the contextually relevant features. This is surprisingly plausible in many cases, as illustrated by some of the more difficult examples presented above. But what if no such rules or facts exist? In that case, even though the hypothesis will specify the appropriate representational vocabulary, standard indexing techniques will not be able to retrieve any applicable rules or facts on the basis of that information. We are faced then with the following dilemma: Either we must abandon the attempt to perform the necessary inferential processing in a directed fashion by using non-inferential techniques to narrow the set of potentially useful rules in such cases, or we must employ non-standard, and more powerful, methods of indexing and retrieval in order to select the appropriate rules.

This is exactly the sort of problem that has motivated renewed interest in parallel models of "spreading activation" or "marker passing" in inference and language processing (see, e.g., Fahlman, 1979; Charniak, 1983). Unlimited parallel inference, *per se*, requires a highly interconnected network of enormous numbers of very powerful processors, and therefore seems unattainable, either in computers or human brains. The basic idea behind these

approaches, therefore, is to see what can be accomplished using much simpler processors, something which will be less than full-blown inference, but which will still be useful (Minsky, 1968). Fahlman (1979), for example, represents an attempt to use marker passing in a semantic network to perform a certain limited class of inferences. Charniak (1983) argues that it should be used to perform a kind of parallel pseudo-inference, the results of which must be checked over and embellished by a more standard serial inference process. In particular, he proposes that markers should be propagated from the concepts mentioned in a text to find paths that connect them in semantic memory by intersection search (as also proposed in Quillian, 1968). Such a path, on Charniak's view, should constitute a potential explanation of the input concepts -- or rather, a set of implication rules which, if they can be properly applied, will constitute such an explanation. Proper application of the rules entails finding appropriate variable bindings, and in particular, determining whether or not *all* of the conditions necessary to apply the rules hold true.

From the perspective of an integrated approach to explanation-based understanding, the important thing about intersection search by marker passing or spreading activation is that it is more powerful than traditional indexing techniques. Using such methods, it may be possible to retrieve inference rules that are likely to be useful in translating the input into terms commensurate with the hypothesis even when no single rule does so directly. For example, suppose our original story ended as follows: "Wesson decided to call some people he knew in politics." One explanation of this action in terms of the hypothesis of revenge is that Wesson is lobbying to elicit support from potential allies in his attempt to block Smith's election to Congress.

Undoubtedly, there exists a counter-planning rule to the effect that when engaged in a struggle against another agent, it will be useful to gather allies. The problem is that although all of the elements in the above input are consistent with the use of this method, each of them taken alone is only a weak indicator. In order to elicit allies, it is necessary to communicate with them, but communication plays a role in almost any counter-planning method. It is easier to get in touch with people who are already acquaintances, and they are possibly more likely to respond favorably to such a proposal. But suppose the input were simply that Wesson decided to call some friends. In this case, it would seem just as likely that he was seeking advice, or simply wanted to blow off steam. Thus, it also seems important that people he is calling are already involved in politics, and are therefore more likely to be useful as allies in a political struggle.

In sum, the problem posed by this input is that each factor, taken on its own, does not

seem sufficiently specific to indicate the particular relevance of any one counter-planning rule over the others, while all of them taken together seem capable of doing so. Of course, if the understander simply attempted to apply all of its counter-planning knowledge to this input, one rule would turn out to fit better than the rest, but our goal here is to attempt to find some less expensive technique to direct the more expensive process of actually attempting to apply such implication rules. If certain combinations of factors could always be guaranteed to be present, then discrimination trees might be an appropriate indexing technique. However, such a guarantee seems to be exactly what is lacking in cases like the above. In such cases, then, marker passing might be an appropriate indexing technique for retrieving relevant explanatory rules associated with the hypothesis. In part II, however, we will see that there are many situations in which marker passing and spreading activation techniques cannot be applied.

10. Conclusion

It is by now generally accepted that understanding a situation entails inferring the hypothesis that best explains that situation. But why do such explanations matter? What does an understander gain from inferring an explanation for some situation? Consider one of the most widely accepted views of *scientific* explanation: Explanations, theories, and hypotheses are useful because they provide *predictions* -- or, to use the term we have been using here, *expectations*. The utility of expectations derived from an hypothesis about some situation should be clear, if not for understanding then certainly for planning. For example, if you block another agent's goal, a goal which is of some importance to him, then it is probably a good idea to watch out for the possibility that he may attempt to gain revenge. In the scenario we have been considering above, it would do Smith little good to lose the election, and then realize afterwards that Wesson had played a role in his defeat. It would be far better for him if he were aware of this possibility during the election campaign, and took steps to deal with it then. Hypotheses and explanations about a situation are important because of the expectations which they provide a planner while he is in that situation. After the fact, it matters less what the explanation was except insofar as it helps the agent learn to make better predictions in the future.

If it were possible to predict, in exact detail, all of the possible outcomes of a situation, and if it were possible to prepare, in advance, to deal with all of them, then it would barely be necessary for a planner to pay attention to its environment at all. But in most cases, these are not possible. We have neither the knowledge nor the inferential resources necessary to predict in detail all possible outcomes of a situation in which we might find ourselves. Nor do we have the real world resources necessary to prepare for all possible contingencies even if exact

prediction were possible. We must, therefore, make do with rough hypotheses and vague expectations, and we must be able to shift resources to deal with contingencies as they arise.

From the point of view of understanding, the crucial phrase in the above discussion is *"watch out for."* If an agent is to be able to shift resources quickly enough to do some good, he must be able to determine which outcome seems likely before the situation has run its course. In other words, if the expectations generated by an explanatory hypothesis are to be useful in such a context, rough as they are, then it must be possible to identify the situations to which they apply as those situations unfold. Thus, in order to apply the expectations which are the *raison d'etre* of explanatory hypotheses, expectations and inputs must be characterized in commensurate terms. But the number of possible construals of an input situation, the number of features in terms of which it might be characterized, is almost limitless. It seems no more possible to construct all possible construals of an input situation in real time than it is to specify all possible outcomes in exact detail ahead of time. Because it seems computationally infeasible to construct all possible construals of an input in real time, choices must be made. But, if those choices are simply made arbitrarily, the resulting representation is likely to be useless, at least with respect to the timely application of the understander's expectations and hypotheses.

In view of the enormous number of potential characterizations of an input, therefore, the process of determining an appropriate characterization must be an *integrated* process: The context provided by a potential explanation must somehow be taken into account to help determine the appropriate representational vocabulary in terms of which to characterize that input, if the expectations generated by that explanation are to be at all useful. That is, an hypothesis about a situation can only be useful if the understander can use it to help determine the relevant features of that situation. In sum, an understander who forms hypotheses or potential explanations for a situation has an advantage over one who does not. Potential explanations must have some functional utility for an understander. We have given one in this chapter.

PART II

CHAPTER 6

PLANNING AND THE UNEXPECTED

...chance favors only the mind that is prepared. -- Pasteur

1. Introduction

The crucial role played by expectation in cognitive functioning has long been a major theme in artificial intelligence and cognitive science. Expectations focus attention on the most salient features of the situation in which an intelligent agent finds itself. Because any situation can be described in so many different ways, in terms of so many different sets of features, it can be argued that without expectations it would be impossible to construct useful representations at all. For in order to construct a representation of some situation, it is first necessary to decide which features of the situation to attend to -- that is, the set of features in terms of which to describe that situation -- and such a decision, however taken, in and of itself constitutes a form of expectation. Whatever other functions expectations may serve in planning and understanding, fixing the vocabulary in terms of which to represent a situation is probably the most important.

In the simplest cases, the set of features to be expected may be fixed in advance, so that a process would have no options as to how to represent the situation in which it operates. In that case, although we may think of the process as having expectations, those expectations stem solely from decisions made in the design of the process -- in the case of an organism, during its evolution -- or as a result of learning. Such a "fixed set" model of expectation may be appropriate under certain restricted circumstances. If the situations to which a process -- or an organism -- is exposed vary along only a few dimensions, then the behavior of that process (or organism) can be specified without explicitly mentioning those features that describe the unvarying background. Thus, its "perceptual apparatus," if you will, may be similarly constrained without affecting the ability of the process to carry out its function -- or the organism's ability to survive. Similarly, if the goals and plans of the process or organism are themselves fixed and few in number, then only those features that tend to be causally implicated in carrying out those goals and plans must be attended to by that process or organism.

If the restricted circumstances described above do not hold, however, the "fixed set" model of expectation will no longer suffice. Thus, a process with many and varied goals and plans, exposed to a highly varying environment -- in other words, an intelligent organism living in the real world -- must have some control over the set of features in terms of which it will represent a situation in which it finds itself. Features that are unimportant in one context may be crucial in another. Factors that don't matter in the pursuit of one goal may be very important in the pursuit of some other. As the situation changes -- and, especially, as new goals and plans are generated -- the importance of a given feature will vary considerably. Under such circumstances, decisions about which features matter must be mutable. Expectations cannot be fixed: They must change as goals and conditions change.

However, the very complexity of the environment which, in part, necessitates the context-dependent variation of expectations, also gives rise in many cases to situations which confound those expectations. In one sense, this situation underscores the ubiquity of expectations in cognitive processing: The ability to understand that something unexpected has happened inherently presupposes a background of expectations against which the unexpected features stand out. But there is a paradox here. Before the unexpected features can be seen in contrast to the expected, they must somehow be represented. Yet the decision to represent a given feature of a situation in itself constitutes, as I pointed out above, a fundamental and characteristic form of expectation. It seems, then, that in certain ways we must be expecting the unexpected -- in a sense, we are always prepared to be surprised. Without perfect knowledge and infinite inferential capacity, expectations -- however necessary -- can never be perfect.

It would only be logical then that intelligence requires the ability to recognize and deal with the unexpected. But such unexpected situations are not just a burden to an intelligent agent: They offer opportunities as well. Two important questions arise in the attempt to construct models that are capable of dealing with unexpected situations: First, how can an agent recognize that something unexpected has happened? And second, how can an agent exploit an unexpected situation once it arises?

2. Recognizing the unexpected

As I pointed out in the last section, the ability to recognize the unexpected has a certain paradoxical air about it. On the one hand, the ability to even represent an unexpected feature of a situation signifies some degree of expectation. On the other, the degree of expectation cannot be very high, since otherwise the feature would not warrant being labelled as "unexpected."

Furthermore, there are presumably an infinite number of unexpected features of any situation, and only a few will warrant being noticed at all -- which is of course one of the key functional motivations for some notion of expectation in the first place. How then can an agent come to notice that something unexpected has occurred?

In the simplest case, the agent will have an expectation about some situation which is *explicitly* denied by that situation. The unexpected feature then is simply the absence of an expected feature, the presence of a feature which was expected to be absent, or perhaps a "mangled" form of the expected feature (which can be analyzed as the unexpected presence or absence of sub-features). For example, an agent might expect that the performance of some action in a plan will yield a certain outcome, and yet that outcome may fail to materialize. Or, an agent might expect that something disastrous will occur in a given situation, and be pleasantly surprised when it does not. These sorts of *expectation failures* constitute the basis of Schank's (1982) theory of learning and reminding, the crux of which is that the failure of an expectation should remind an agent of similar failures in the past, and that the explanation of such failures can lead to better expectations in the future. Expectation failures are also play a fundamental role in Sussman's (1975) theory of learning as debugging almost right plans. These theories, then, offer one vision of how an agent can exploit an unexpected situation by *learning* from it.

In principle, given that the presence or absence of the appropriate feature can be determined, its unexpected status will be obvious at face value. In practice, however, it is not always easy to determine that something unexpected has happened even in such simple cases, because it is not always clear whether a situation actually denies the presence of an expected feature. For example, consider a language understanding system based on some form of script application, and suppose that some event expected by the script is not mentioned in the text. Is the system entitled to assume that the event did not take place? In most cases, the answer is that it is not: In fact, such systems are supposed to infer that the unmentioned event has taken place. But what works in most cases does not work in all: There will be circumstances under which the failure to mention an expected event actually does signify that the event failed to occur.

It is far more difficult to recognize that something unexpected has occurred when the unexpected feature of a situation cannot even be considered an expectation failure at all, because it is something about which the agent had no expectation one way or the other (see Lalljee and Abelson, 1983, for a discussion of how this distinction affects the process of constructing explanations for anomalous situations). The difficulty posed by such unexpected

features stems from the fact that they really are innumerable in any situation. So the problem facing an agent is to notice unexpected features which are somehow interesting or significant. Yet to ascribe interestingness or significance to a feature requires a motive: The agent must understand *why* the unexpected feature is significant. Here then is where the paradox is felt most acutely: In order to determine whether an unexpected feature is significant, it must first be noticed. Yet an agent cannot notice all unexpected features, there are simply too many in any situation. So, somehow he must come to notice those features which are significant without being swamped by those which are not.

How can an agent come to see that an unexpected feature of a situation is significant, or to understand its significance? In order to answer this question, we must first have a theory of what makes something significant to a given agent. *Significance can only be defined with respect to goals.* The significance of a feature to an agent lies in its relation to that agent's goals. These goals may be very general, e.g., curiosity, or very specific, e.g., the need to get to Milwaukee. Moreover, the relation between a feature and a goal may vary as well: Is the presence of the feature a necessary prerequisite for a plan for the goal? Does it indicate likely success or failure? Does it signal the need for caution or other special measures? Whatever the answers to these questions, the problem of recognizing unexpected situations -- because its solution involves some notion of significance, a notion which can only be defined with respect to goals -- leads us, rather directly, to the subject of planning and plan execution.

3. Unexpected situations and opportunistic planning

As we have seen, the issues raised in dealing with unexpected situations are intimately related, not only to understanding, but to planning as well. An agent's understanding of an unexpected situation depends on his ability to represent that situation as possessing a feature which was unexpected. Since every situation possesses an infinite number of such features, there must be some other characteristic besides unexpectedness that matters. That characteristic can only be significance with respect to the agent's goals and plans. Thus, an agent's understanding of an unexpected aspect of some situation, as much as of an expected feature, depends on his goals and plans.

The above argument suffices to establish that some notion of significance with respect to the goals and plans of an agent is crucial in the agent's understanding of unexpected situations. But, in order to understand this relationship, we must turn our perspective around and ask why it would matter that goals and plans affect the understanding of unexpected situations. In particular, what impact do unexpected situations have on plans and goals, and how must a

planner be prepared to deal with them?

Quite often, perhaps even most of the time, unexpected situations mean trouble in planning. Expectation failures, in particular, are likely to coincide with the failure, or impending failure, of a plan. For example, an action taken within a plan may not have its expected -- and, presumably, desired -- effect, and hence the continuation of the plan may be jeopardized. Or a normally unproblematic precondition, the satisfaction of which is expected either because of prior efforts by the planner or because it has simply been satisfied by the environment in the past, may turn out in some case to be absent. These kinds of unexpected situations, because they stem from the explicit failure of one or more expectations, are relatively easy to detect. Thus, they have received some attention in the planning literature, under the rubric "execution monitoring" (see, e.g., Fikes, Hart, and Nilsson, 1972; Sussman, 1975; Sacerdoti, 1977). When such a failure arises, the planner can attempt to save the plan by a bit of on-the-spot replanning for the subgoals which have failed.

The failure of an expectation stemming from a plan may also signify good news to a planner. One might have an expectation that some particular problem will tend to arise in a certain type of planning situation, and happily find that it does not arise in the current instance of that type of situation. In the simplest cases, some precondition which must usually be planned for, and to the achievement of which some effort must usually be devoted explicitly, will be found to be serendipitously satisfied in the current situation. Most theories of planning can handle this situation, in the rather minimal sense that the planner will use the satisfied precondition rather than redundantly working to establish it again.

In both harmful and beneficial expectation failures, however, simply dealing with the expectation failure on the spot is not enough. A planner should, in the best case, come to understand *why* the expectation has failed. Why is the formerly unproblematic precondition now problematic? Why is the formerly problematic one now unproblematic? Some work has been done on this problem in the case of a harmful failure, perhaps because the reasoning required to understand the source of the failure is often necessary simply in order to patch the plan and continue. That is, the explanation of harmful failures is driven by the needs of effective planning and plan execution. Thus, even though learning may result from the explanation of such failures, it is not necessary to invoke learning as a goal to motivate the explanation process (although see Sussman, 1975, and Hammond, 1986, for theories of learning based on the explanation of plan failures). However, the sort of expectation failure in which something which was expected to be a problem turns out not to be, quite obviously does not pose the same difficulties for plan execution. That is, an agent need not explain such an

expectation failure in order to successfully complete the plan. It need only note its good fortune and continue on. Such cases therefore require invoking the need to learn in order to justify the need to explain the failure.

However, although the importance of expectation failures is best explained in terms of plans and goals, this may not be necessary given the prior existence of the expectations themselves -- although, admittedly, it does seem difficult to justify the existence of expectations in a system without goals -- and this possibility forms the basis of learning theories which are driven by expectation failures and yet take no account of the learner's goals (see Schank, Collins, and Hunter, in press, for a critique of such theories). The importance of goals and plans in the understanding of unexpected situations is, however, indisputable in the case of truly unexpected features of a situation, i.e., those which neither confirm nor deny an expectation. Here, the problem of somehow discerning significant unexpected features of the situation -- given that any situation can be described in terms of a myriad number of features, most of which are unexpected, but most of which simply don't seem very important -- cannot be avoided. The most intuitively plausible and functionally sensible way to define significance is in terms of an agent's goals. Thus, significant features of a situation would be those which somehow had an impact, in either a positive or negative fashion, on the agent's plans and goals.

Let's consider how this notion might be applied to the problem of distinguishing significant unexpected features of a situation. The features are *significant* because they affect the agent's plans and goals. They are *unexpected* because they either confound an expectation -- i.e., they are expectation failures -- or because they are truly unexpected -- i.e., they neither confirm nor deny an expectation. Consider again, briefly, the case of explicit expectation failures which are significant. Since they are significant, they must somehow affect the plans and goals of the agent. The plans and goals which they affect, of course, are the plans and goals which gave rise to the violated expectations in the first place. But under what circumstances would plans and goals give rise to expectations? Most likely when they are actively being pursued. In sum, we are led to the not very surprising conclusion that the features involved in expectation failures are significant with respect to the goals and plans which were being actively pursued and which, for that reason, gave rise to the expectations that were violated.

The more interesting question then is what goal or plan a *truly* unexpected feature affects if it is significant. If the goal or plan were currently under active pursuit, then it would presumably give rise to expectations about significant features, i.e., features which would be

likely to affect it in either a positive or negative way. Thus, if a truly unexpected feature is significant with respect to a goal or plan under active pursuit, then it must be because the agent's knowledge of the features which are important to that goal or plan is incomplete in some way. But in that case, it is highly probable that the agent will eventually experience an expectation failure of the more direct kind, e.g., his plan is likely to fail. Thus, before an agent will detect that a truly unexpected feature has affected his current goals and plans, he will likely notice that an expected feature has been confounded. The relevance of the truly unexpected feature will then be due to its causal role in that expectation failure. Thus, if a truly unexpected feature is noticed because it is significant with respect to a goal under current pursuit, then it will most likely be discovered in the course of explaining the failure of an expectation stemming from that goal.

There is, however, an alternative possibility. A truly unexpected feature need not necessarily be significant because of some previously unknown relationship with a goal under active pursuit. It could also be significant by virtue of a relationship -- either previously known or unknown -- to *other* goals or plans of his, goals and plans which are *not* under current pursuit. But why would it be useful for an agent to be able to recognize such features at all? This question can only be answered by considering what an agent would be able to do if it could recognize a feature relevant to a goal other than the one it was actively pursuing that it could not do without that ability.

Some of the advantages are obvious. Consider a person walking across the street, with the goal of getting to a particular restaurant in order to have lunch, and talking to his lunch companion. If a car suddenly careens around the corner and starts accelerating towards him and his friend, then that is an unexpected situation which is relevant to a goal -- staying alive -- other than the ones he was actively pursuing, namely going to lunch and engaging in conversation. Obviously, then, a person must be able to recognize the unexpected feature -- the car speeding towards him -- and why it is significant -- that he or his companion could be hurt or killed. The real-time ability to detect and avoid unpredicted -- and in fact unpredictable -- dangers is thus one that requires the ability to notice features of situations which are relevant to goals other than those actively governing the agent's current behavior.

There are other ways in which unexpected features can relate to goals besides endangering them, however. An unexpected feature can be significant, not because it poses an unexpected threat, but because it poses an unexpected *opportunity* -- that is, it facilitates the pursuit of some goal other than one which is currently governing the agent's behavior. Such a feature might be the unexpected presence of some highly problematic precondition of a plan for

the goal. If that precondition were only intermittently available, then it would be vital to notice its presence on those occasions when it were in fact present. For example, suppose that you want to buy some item, but the price is too high for your budget. If, in the course of pursuing other goals and plans -- e.g., reading the newspaper, or shopping for some other item, or watching television -- you find that the item is on sale, and the price has been reduced to the point where you can afford it, then it would be useful to recognize that fact. Thus, in order to recognize and seize opportunities, an agent must be able to notice unexpected features and recognize their significance with respect to goals not necessarily under active pursuit.

The ability to recognize and seize opportunities has implications for both the process by which goals are set -- goal activation -- and the process by which plans for those goals are constructed. In most models of planning and plan execution, goals are activated in order to serve as subgoals to previously active, higher-level goals. The top-level goals which start the ball rolling are simply a given, activated at the start of the process. In an opportunistic planner, in contrast, goals can be activated not only in the service of previously active, higher-level goals, but also by the presence of opportunities for their pursuit. In addition to its impact on goal activation, opportunistic processing can also affect the manner in which plans are constructed. (This idea is originally due to Hayes-Roth and Hayes-Roth, 1979, as is the term "opportunistic planning.") That is, when planning to achieve several goals, a planner may recognize an opportunity to achieve one of those goals rather easily in the course of constructing a plan to achieve another of those goals, and alter his plan accordingly.

However, although the application of opportunistic processing to goal activation, on the one hand, and to plan construction, on the other, can be pursued separately, in fact the two have a great deal to do with each other. An agent who is capable of opportunistic goal activation may, while constructing a plan to achieve one goal, recognize that his plan offers the opportunity to achieve some other goal as well, even though he had no prior intention of planning for that other goal at the time. In other words, one goal may be activated in the course of planning to achieve some other goal, and alter the course of planning as a result. Thus, opportunistic goal activation substantially broadens the scope of opportunistic plan construction. Similarly, the ability to construct plans opportunistically is crucial to the ability to successfully pursue a goal which has been opportunistically activated, for once such a goal has been activated, an agent must be able to alter his current plans in order to pursue it. In a minimal sense, that could be accomplished by dropping the current plan and constructing a new one. But it is obviously far more economical alter the plan in a less drastic fashion if possible, and the abilities which are necessary to determine how that might be accomplished are exactly those which are necessary for opportunistic plan construction as well.

4. Integrated processing and opportunistic planning

In chapter one of this thesis, I presented the functional argument for integrated processing in planning and understanding. The fundamental point is that, to avoid backtracking as much as possible, programs should make decisions about which lines of inquiry to pursue on rational grounds, rather than arbitrarily. Because making rational decisions means making informed decisions, therefore, integrated processing entails the use of as much relevant information as seems necessary in making such decisions. As we saw in part I, this principle leads to a relatively top-down conception of understanding. In planning, however, it leads to something quite different.

The problem of choice arises in planning in part because there are usually many alternative plans for achieving a given goal. Moreover, such choices interact with each other. As a result, a great deal of computation may be necessary in order to produce a plan which simply meets the minimum logical requirement of being self-consistent, regardless of whether or not it can be successfully pursued under the circumstances that actually hold. Thus, it seems clear that external factors which militate for or against the success of a given plan should be taken into account as early in the planning process as possible, in order to avoid the high cost of producing plans which are coherent, but must nevertheless be discarded because they happen not to be feasible in the current or expected situation. Indeed, by favoring or forbidding the use of certain plans, the early application of contextual constraints might narrow the set of available choices sufficiently that, if the constraint of constructing an internally coherent plan can be met at all under the circumstances that hold, it can be met without engaging in a great deal of undirected search. Just as in understanding, therefore, taking context into account does not make the problem harder, it makes it easier. It may make our theories more complicated, but that is beside the point.

Thus, in order to avoid arbitrary choice and backtracking, an integrated model of planning must attempt to take external contextual factors into account as early as possible in the planning process. But, as we have seen in this chapter, planning and plan execution in the context of unexpected situations, and in particular the ability to recognize and seize opportunities, similarly entail the joint consideration of both external factors and internal goals. It might be argued that this is hardly surprising: After all, it is tautological that plan execution must take external circumstances into account. But it is important to realize that, were it not for unexpected occurrences, plan execution would be entirely trivial. The point here is that an integrated approach to planning must attempt to reduce arbitrary search by attending simultaneously to both the agent's goals and to the situation in which he finds himself, and this

is also what is necessary for opportunistic processing. Thus, an integrated approach to planning seems to lead towards models in which planning and plan execution are themselves highly integrated and tightly interwoven in order to meet both of these requirements. In the next chapter, we will pursue this theme in the context of a specific problem: planning how to respond in an argument.

CHAPTER 7

ARGUMENTATION: A CASE STUDY IN OPPORTUNISTIC PLANNING

1. Introduction

Engaging in an argument is a complex task of natural language processing that involves understanding an opponent's utterances, discovering what his "point" is, determining whether his claims are believable, and fashioning a coherent rebuttal. Accomplishing these tasks requires the coordination of many different abilities and many different kinds of knowledge: memory and inference, planning and plan understanding, knowledge of the topic under discussion, and knowledge of the structure of arguments and of effective argument strategies. Because arguing, and conversation generally, involve real-time interaction with another agent, this coordination must be even more flexible than is required for other natural language processing tasks. An arguer must have some expectations about what his opponent might say, but must also be able to respond to the unexpected. He must have some idea of the claims he wants to make, and plans for putting them forward, but his opponent may confound these plans. Or, more positively, his opponent may say something that offers an unforeseen opportunity to make a point. Arguing thus exemplifies the need for the flexible integration of top-down and bottom-up processing in both language understanding and production.

This chapter is concerned with the roles of memory processing and planning in the processes of understanding and generating utterances in an argument or conversation. In particular, I will show that the memory and inferential processing necessary in order to understand another person's utterances in an argument or conversation can and should play a large role in generating a response, performing functions that most previous theories of conversation would delegate to explicit, goal-directed planning. This chapter does not present a detailed model of the process of engaging in an argument, or of the representations that such a model would manipulate (although see Birnbaum, in preparation, for a theory of such representations). Rather, my goal is to show that there are certain properties which any such model must have, and to sketch out a general approach to planning which fulfills those requirements.

2. The problem of choice in conversation

One of the most interesting, difficult, and yet frequently neglected questions that arises in analyzing a conversation or an argument is determining why a participant responds to a given utterance as he does, rather than in other equally plausible ways. For example, in the following mock argument between an Arab and and Israeli concerning Middle East affairs, each response [2] through [6] below can be replaced by other responses that, while quite different, are still coherent in the context of the argument up to that point:

[1] Arab: Israel is trying to take over the Middle East.

[2] Israeli: If that were our goal, we wouldn't have given back Sinai to the Egyptians.

[2'] Israeli: No, it's the Arabs who are trying to take over Israel.

[3] Arab: But you haven't given the West Bank back to the Palestinians.

[3'] Arab: You only returned the Sinai because of U.S. pressure.

[4] Israeli: Israel can't negotiate with the PLO because they don't even recognize Israel's right to exist.

[4'] *Israeli: Israel can't allow a hostile state in such close proximity.*

[5] Arab: Israel doesn't recognize the PLO either.

[5'] Arab: How can they? That's their only bargaining chip.

[6] Israeli: The PLO is just a bunch of terrorists.

[6'] Israeli: The PLO has to show its good faith first.

One of the reasons why this problem has not received a great deal of attention is that it need only be directly confronted by a *functional* theory of conversation, one that attempts to assert and to justify claims about the computational processes by which a conversationalist can or should determine how to respond, given his purposes in the conversation. In a descriptive theory, on the other hand, the problem of why a conversationalist responds one way rather than another need not be directly addressed, because the goals of such a theory would be satisfied if it were able to properly delineate the range of possible responses to an utterance, whatever the domain of discussion.

Suppose, for example, that you make a point in an argument, and your opponent immediately attacks that point. A descriptive theory of argumentation would supply a list of your response options. A first attempt at such a list might look something like this:

Attack the input.
Re-support your previous point.
Go back to the main point dominating this exchange.
Change the subject.

In a computational theory, however, such a list can at most be only a first step towards solving the problem. Even if we characterize the conversationalist's task as involving the selection of a response rule from among a set of domain-independent alternatives like that above -- an assumption which is, as I pointed out in chapter one, entirely unjustified, and quite likely to be incorrect -- a computational theory cannot sidestep the question, *how is the choice to be made?* One could, of course, write a program that chose randomly, but that is clearly an evasion rather than a solution. Certainly, it seems implausible that people choose randomly, although they might in some rare circumstances. One might attempt to reduce the magnitude of this problem by better specifying the conditions under which the various options apply, in terms of domain-independent features of the argument context that differentially affect the plausibility of different responses, and this is indeed possible (see, e.g., Birnbaum, 1982). However, even the finest domain-independent categorization of conversational options would be unlikely to completely eliminate the problem of choice, because such a categorization would *ignore two essential and highly domain-specific elements: the content of the conversation or* argument, and the knowledge and goals of the speaker.

Within the framework of an integrated theory, a radically different sort of explanation can be pursued by attending to these other elements. A program capable of engaging in a conversation or argument needs a great deal more than just knowledge about conversations. As with other complex cognitive functions, the explanation for much of the observed behavior can be expected to lie in the interactions and relations among the many different abilities and sources of knowledge needed to converse or argue. The most important of these is an inferential memory capable of organizing knowledge about the world and applying that knowledge in order to understand or generate utterances. Such a memory can be expected to play a key role in explaining conversational behavior.

3. The role of memory processing

It is by now a truism that memory and inference are central elements in natural language processing. Any viable theory of language comprehension must include processing whereby the linguistic input, or some intermediate representation of its meaning, is somehow related to memory, appropriate inferences are drawn, and the input and inferences are indexed and stored

for future use. In light of the complexity of this processing, it seems likely that much of the explanation for conversational behavior arises from the contents and function of such a memory (Schank, 1977). In particular, memory plays a key role in explaining why a participant in a conversation or argument responds as he does, rather than in some other way. This claim is based on the observation that a good response to an utterance in a conversation can often be discovered as a side-effect of the memory and inferential processing that is required simply in order to understand that utterance (Birnbaum, Flowers, and McGuire, 1980; this generalizes Lehnert's, 1979, point that the memory processing needed to properly interpret a question about a previously understood story often leads directly to the answer). Such opportunities greatly simplify the problem of deciding how to respond. Furthermore, the reason why a conversationalist says one thing rather than another can then be explained in part by what he knows about the topic under discussion. Different responses reflect the idiosyncratic states of different people's knowledge and memory organization.

For example, consider again the following exchange in a mock argument between an Arab and an Israeli over Middle East affairs:

[1] Arab: Israel is trying to take over the Middle East.

[2] Israeli: If that were our goal, we wouldn't have given back Sinai to the Egyptians.

The Israeli's understanding of the Arab's assertion [1] involves the retrieval and application of the concept of imperialism, the creation of an instance of the concept with Israel as the actor and the Middle East as the target, and the recognition that this assertion is intended as an accusation. The concept of imperialism is represented by a complex knowledge structure (let's call it **TAKE OVER**), consisting of several component substructures. I will assume that it contains, roughly, the following: the actor **BUILDS UP MILITARY STRENGTH**, **ATTACKS** the target, **CONQUERS TERRITORY** of the target, and then **OCCUPIES TERRITORY** of the target. Given the theoretical framework in which inferential memory processing forms the basis of understanding, the Israeli must relate this knowledge structure to his long-term memory in order to understand input [1]. Thus, for example, upon relating the **BUILD UP MILITARY STRENGTH** substructure to the relevant structures in his memory, he might discover his belief that Israel has indeed engaged in building up its military strength (although in his memory this would be explained by the goal of self-defense). More to the point here is what he might notice in the course of relating the **OCCUPY TERRITORY** substructure to memory. He will, presumably, discover his belief that Israel is in fact occupying Arab territory. But he might well also discover an instance of Israel *relinquishing* conquered territory -- the Sinai -- which contradicts the original allegation of

imperialism. It is this fact, plausibly discovered in the course of understanding the allegation, which forms the basis of the Israeli's rebuttal [2].

That some processing of this kind, involving roughly the memory structures indicated, does in fact occur naturally in the course of understanding can be further justified by considering a scenario in which some country is accused of imperialism, even though it has never engaged in any of the actions that constitute **TAKE OVER**: has not built up its military strength, has not attacked, conquered, or occupied any foreign territory. For example, suppose someone claims to you that Denmark is trying to take over Norway. The fact that Denmark has not made any aggressive moves towards Norway would, I maintain, cause you to be puzzled by this claim, even in the absence of any disposition on your part to argue against it. But noticing that Denmark has not made any aggressive moves towards Norway would entail exactly the sort of processing described above for utterance [1]. Thus, such processing would seem to occur independently of any intention to argue against the input.

As another example of this kind of processing, consider the following continuation of the previous exchange:

[3] Arab: But you haven't given back the West Bank to the Palestinians.

Both the Israeli utterance [2] and the Arab response [3] refer to Arab territory occupied by the Israelis. It seems entirely reasonable to suppose that this topic is sufficiently important to an informed supporter of the Arab position to warrant the existence in his memory of some knowledge structures which organize information relevant to it. In particular, these knowledge structures would point to instances of **OCCUPY TERRITORY** which have Israel as the actor and former Arab lands as the target. Further, these would be the exact structures which could be expected to play a role in the inferential memory processing needed to understand utterance [2]. Thus, in the course of trying to understand the utterance, the Arab would naturally be reminded of instances of continued Israeli occupation of Arab territory. One of these instances -- the Israeli occupation of the West Bank -- forms the basis of the Arab response [3].

The main point to be made here is that the memory processing that uncovers these rebuttals is not particularly argument-driven -- although the decision as to whether or not to use them undoubtedly is. It is, more or less, the kind of processing that would be necessary whether the utterance occurred in an argument or in some other context. Indeed, there is a large class of inputs for which this is clearly true, namely those cases in which an input includes a factual claim that in some way directly contradicts what the understander believes.

For example, suppose that you were arguing with someone about the Viet Nam War, and he claimed that the Communists were responsible, because they refused to participate in the U.N.-sponsored elections intended to settle the political future of the country following the French pull-out. This claim happens to be false: It was the government in the South, which, at the urging of the U.S., refused to participate in the elections. If you knew that fact, then you would undoubtedly notice that your opponent's claim was incorrect, and you would certainly say so in the argument. The point here is that you would probably also notice that the claim is incorrect if you encountered it in a context other than an argument, for example, if you happened to read it in a newspaper account of a Presidential press conference.

The conclusion to be drawn is that the memory processing needed to understand and assimilate an input must, as a matter of course, notice contradictions or inconsistencies -- among other relations -- between that input and the relevant beliefs of the understander. It seems obvious that if someone tells you something that you believe not to be the case, and you understand what he is saying, then you will notice the contradiction with your beliefs. This is true regardless of whether or not you are engaged in an argument at the time. Nor is it limited to situations with emotional or ideological overtones. Suppose that, in your presence, someone were giving directions to a third party, and he incorrectly said to turn right at some point, when it should be left. If you had been paying attention to the directions, then you would undoubtedly notice that error, and you would probably say so.

The ability to notice contradictions between new information and old, and its importance, have been the subject of a great deal of research on memory processing. Rieger (1975) specifically referred to this ability as one of the functional justifications for his theory that understanding involves a great deal of undirected inference. Fahlman (1979) specifically addressed the problem of noticing certain straightforward kinds of contradictions, which he termed "clash detection." More recently, Schank (1982) addressed the more general question of what kinds of relations memory must be able to perceive between an input and prior knowledge, by investigating how and why people are reminded of a previous experience or story when attempting to understand a new one. Expectation failures -- contradictions between what is expected and what in fact occurs -- are one of the chief driving forces in learning and memory organization in his theory.

The utility of this sort of memory processing in an argument should be obvious. If a contradiction is noticed in the course of understanding an opponent's utterance, then that contradiction is a good candidate to form the basis of a rebuttal. The same general approach also points to a possible explanation of topic shifts. In the course of understanding an

utterance, a conversationalist might be reminded of something related, although perhaps incidental, that he feels is interesting or important enough to bring up. In general then, an increased reliance on memory allows the content of the discourse, and the speaker's knowledge, to play more active roles in the formation of a response.

4. The role of top-down planning

Just as important in determining a speaker's response as his knowledge or the content of the discourse are his goals in the conversation or argument, and the plans by which those goals can be achieved. Indeed, much recent research on conversation has been based on the idea that conversation and other forms of discourse are planned behavior, in the same way as most other intelligent action is (see, e.g., Levin and Moore, 1977; Deese, 1978; Grosz, 1979; Hobbs, 1979; Levy, 1979; Allen and Perrault, 1980). Most prior investigations of conversational behavior have employed some notion of planning borrowed from the problem-solving literature. One of the most influential genealogies in this literature -- and certainly the work that has had the most impact on theories of conversational planning -- has been the line starting with GPS (Newell and Simon, 1963), and continuing on through STRIPS (Fikes and Nilsson, 1971), and NOAH (Sacerdoti, 1977). All of these planning models are top-down in the rather straightforward sense that they all start with an explicit goal, and, with varying degrees of sophistication, attempt to devise a plan, or sequence of actions, that will satisfy that goal. NOAH is top-down in another way as well: It uses a method of progressive refinement, starting with a general, abstract plan, and gradually expanding it to a level of detail which can be executed.

The most obvious reason why it is necessary to plan top-down in order to carry on a conversation is that it seems implausible that a good response to an utterance can always be found in the course of understanding it. Perhaps the simplest example that such top-down conversational planning is possible is *lying*. Sometimes when a person finds himself with no good response in an argument or conversation, he may lie by fabricating a fact, or a quote, or by embracing a position that he does not really believe. In order to construct a lie, however, one must decide what kind of lie would be useful, i.e., what goal the lie should serve. For example, in an argument a lie might serve to buttress a claim of the speaker's that has just been attacked, or to attack some of the evidence supplied by his opponent, or even to attack his opponent's character. Choosing one of these goals, and then constructing a lie that achieves it, is a clear example of top-down planning in a conversation.

An exclusively top-down approach to planning can work in situations which are more or

less under the control of the planner. Thus, it has proven useful in generating single utterances in a cooperative situation, as exemplified by the work on speech acts (Cohen and Perrault, 1979), or in conversations about tasks which themselves have a planned, hierarchical structure. But conversations and arguments do not, in general, meet those requirements. The actions of one's conversational partner, who in the case of an argument is assuredly not disposed to cooperate, can be expected to interfere with any top-down plan spanning several exchanges. Utterances in a conversation must not only further the speaker's own goals; they must also relate to what his partner (or adversary) has just said. Thus, unless a speaker can predict, rather specifically, how his adversary will respond, his utterances cannot be completely planned in advance.

Approaches to response generation based on the primacy of domain-independent argument or conversational mechanisms, whether conceived of as planning or otherwise, also suffer the drawback of being overly top-down. Such approaches typically produce a response as the result of a series of hierarchically arranged choices which implement, rather directly, a descriptive theory of the sort discussed earlier in the chapter. (For example, Reichman, 1981, describes a fairly extensive model along these lines.) The problem they face stems from the difficulty of choosing among alternatives in the absence of any reason to believe that the choices will actually lead to a good response. For example, at the top level, the choices might be as follows:

Attack one of the opponent's claims.
Re-support one of your own claims.
Change the subject.

Suppose that the attack option were chosen. The next level of choices might then be concerned with deciding which of the opponent's claims to attack:

Attack the last thing he said.
Attack the claim that his latest utterance supports.
Attack his chief claim dominating this exchange.

Suppose now that the first choice were taken. The next choice would be a decision as to what kind of attack to make, and the options would to some extent be dependent on the nature of the utterance to be attacked:

Attack the relevance of the claim.
Attack the truth of the claim.
Attack the authority cited.

Now suppose any one of these were chosen. In order to formulate a good response based on the decision to attack the relevance, truth, or authority of some claim, some evidence must be found in memory to show that the claim is irrelevant or untrue, or that the authority is arguable. The problem is, there is no reason to believe that any such evidence actually exists, and the utility of the entire chain of decisions is in jeopardy if none is found.

Thus, domain-independent approaches to response generation are too top-down, because too many decisions have to be made in the absence of any knowledge about what, ultimately, will be needed in the way of facts and reasoning about the topic domain to carry out a response based on those decisions. Such approaches must therefore rely heavily on back-up in order to produce useful plans. As I have argued repeatedly, the use of arbitrary choice and backtracking must be avoided in a genuine process model if at all possible. Thus, in a genuine process model of conversation, decisions about what to say must take into account, as early as possible, what is known about the subject under discussion. The exploitation of opportunities for rebuttal uncovered while trying to understand an input is one method for producing a response that is consistent with this principle.

5. Opportunistic processing: A synthesis

The conclusion to be drawn from the examples and discussion of the last four sections is that participating in a conversation or argument requires opportunistic processing, that is, both the ability to formulate plans, and the ability to recognize and pursue opportunities, in order to satisfy conversational goals (McGuire, Birnbaum, and Flowers, 1981). Obviously, the need for this kind of flexibility is not limited to conversations: It is a key factor in all kinds of intelligent behavior. To take an example from Meehan's (1979) TALE-SPIN domain, suppose that Joe Bear is hungry, and decides to ask Wilma Canary where some honey is. She tells Joe that she will answer him, if he brings her a worm. In the course of looking for a worm, Joe stumbles across some honey, or perhaps some fruit. If Joe does not eat the honey or the fruit, but rather continues searching for a worm to give Wilma, we would say that his behavior was not very intelligent. But without the ability to notice opportunities in the world that can be used to satisfy goals other than the one which is immediately governing his current behavior, that is exactly what would happen.

Recent research on planning and problem-solving has begun to address this point. Hayes-Roth and Hayes-Roth (1979) have proposed a model of opportunistic planning in which the planner's decisions in formulating a plan are not strictly hierarchical. Rather, decisions and observations at a given level of abstraction in the plan can influence not only the more specific levels that it dominates, but can also suggest opportunities to more abstract levels. Thus, for example, when planning to buy several items, including milk and eggs, the planner will notice an opportunity to achieve one of its currently active goals -- to buy eggs -- while constructing a plan to go to the store to buy milk. It will then plan to buy both milk and eggs at the store. (Hobbs, 1979, has pointed out the influence of something akin to this form of opportunism in conversational behavior.) However, their model only exploits opportunities that arise while planning, not while executing plans. Thus, new goals cannot be formed as a result of noticing opportunities to achieve them. The difference between these two varieties of opportunism can be illustrated with another story from the shopping domain. Suppose a planner decides to go to the store to buy milk. In the course of executing that plan, while at the store, he notices that eggs are on sale. Realizing that he will need eggs in the near future, he checks to see whether he has sufficient funds, and if he does, he buys some.

To take another example from the TALE-SPIN domain, suppose Joe Bear is searching for a worm to give Wilma Canary, so that she will tell him where he can find some honey. If in the course of that search he should happen to come across some water, then, if he is thirsty, he should realize that fact and drink some water. It is this sort of execution-time opportunism that is necessary for Joe Bear to behave intelligently in the TALE-SPIN vignette presented earlier as well.

Wilensky's (1983) theory of *meta planning* comes closest to satisfying the requirements for full opportunistic behavior. The most salient feature of the theory is its emphasis on goal detection as a central issue in planning and problem solving. That is, a planner must be able to figure out what its goals should be in a given situation. In Wilensky's model, the chief application of this ability is in noticing some interaction between several active goals -- for example, that two goals are in conflict -- and as a result formulating a new goal to deal with that interaction. However, the ability to determine what goals are relevant is crucial to opportunism as well. In order to exploit an opportunity, one must first recognize that it is an opportunity, and to do that, one must realize that the opportunity serves some goal that might be worth achieving. A goal detection mechanism must therefore be able to suggest relevant goals upon the detection not only of problems, but also of [illegible] opportunities.

After an opportunity is noticed, the next steps [illegible]

the opportunity. In actuality, this must be a decision as to whether or not to pursue the goal that the opportunity can further. This in turn depends on what other goals the planner has active, including both previously active goals and those presented by other new opportunities, and of course, the time and resources that will be necessary in order to achieve each of these goals. In other words, much of the effort required for opportunistic planning must be devoted to reasoning about goals: their desirability, their cost, and their interactions. Reasoning about goal interactions, in turn, requires reasoning about the possible plans for carrying out those goals, and their interactions.

At this point, the complexity involved in opportunistic processing might seem prohibitive, especially in a time-limited situation like a conversation. One way to reduce this complexity is to pre-plan, not at the level of what actions to perform given a certain situation in the world, but rather more abstractly about what goals to pursue in a given sit ıation characterized by several active and interacting goals. For example, if one has some idea of the kinds of goals that are likely to arise in some situation, setting their relative priorities ahead of time, instead of having to figure them out on the fly, is probably very worthwhile. Or, one might plan to always choose, from within a given class of desirable goals, the cheapest opportunity that presents itself. This sort of pre-planning, at the level of priorities and costs, seems to correspond to our intuitive notion of the strategic level in planning.

The application of these ideas to planning in arguments requires identifying argument goals, and strategies that choose among several possible goals. A very simple example of such a strategy might be to always exploit an opportunity to attack an erroneous factual claim of your opponent's. This strategy would be cheap to use, because most such opportunities would arise during the attempt to understand the erroneous claim. Thus, the work of determining that the claim was false, and why, would already have been accomplished. Because it is cheap, it would easily satisfy the general imperative of responding to your opponent's last utterance in a relevant way. If effectively pursued, the strategy would satisfy the goal of casting doubt on your opponent's credibility, and adding to your own. In this respect the strategy is quite aggressive. But in another sense, the strategy is rather passive: It would allow your opponent to set the agenda by making provocative claims. Thus, the unconditional use of this rule would reflect a decision, whether explicit or not, to set a higher priority on the goal of attacking the opponent than on the goal of controlling the agenda.

6. Conclusion

At the end of the last chapter, I argued that planning decisions must take external context

into account as early as possible in order to avoid backtracking. I also pointed out that the ability to take such external considerations into account during planning and plan execution was necessary in order to recognize and seize fortuitous opportunities to pursue a goal. In other words, I argued that an integrated model of planning would also be, in the best case, an opportunistic model of planning. In this chapter, I have shown that these two requirements do in fact converge in the domain of conversational planning, particularly in arguments: A conversational planner which is capable of avoiding arbitrary choice and backtracking in constructing a response will be an opportunistic conversationalist.

An opportunistic planner must set its goals not only as a result of planning to achieve previously active, higher-level goals, but also by assessing the goals that its current situation presents opportunities to pursue. However, the ability to recognize when a situation seems to facilitate the achievement of a worthwhile goal puts a heavy burden on the perceptual, inferential, and memory capabilities a planner uses to understand and assess the situation in which it finds itself. To some extent, this burden can be eased if the planner has a general characterization of the sorts of opportunities that might arise in a given situation. For example, it seems reasonable that, when going to a grocery store, a planner should expect that some items may be on sale, although exactly which items will be unknown. However, in order to notice truly unexpected opportunities, a planner must be able to infer new goals from features of the situation not necessarily related to its currently active goals. We will return this problem in the next chapter.

Finally, the application of these ideas to conversational behavior offers the promise of a model that avoids many of the problems inherent in previous approaches. An opportunistic conversationalist would be able to set conversational goals in part on the basis of what its inferential processing uncovered in the course of understanding the content of the discourse. In particular, this would lead to a less top-down approach to the formation of responses, since the discovery of a potential response would determine which conversational goal to pursue as often as, if not more often than, the reverse. It is also worth noting here that the situations which provide these opportunities for response are themselves the result of the inferential memory processing that constitutes understanding: Opportunities exist not just in the world but in our thoughts.

CHAPTER 8

RECOGNIZING OPPORTUNITIES

1. Introduction

The most difficult and important component of opportunistic behavior is simply *noticing* that there is an opportunity, for the ability to exploit an opportunity depends, first and foremost, on being able to recognize its presence. An opportunity exists when the situation in which an agent finds himself meets certain conditions that may facilitate the pursuit of one of his goals. Thus, there are two aspects to the problem of recognizing an opportunity. First, an agent must be able to recognize the presence of those features of the situation that constitute the opportunity -- that is, the conditions that facilitate the pursuit of some goal. Second, he be able to recognize that those features do in fact constitute an opportunity -- that is, he must be able to determine the goal for which those features constitute an opportunity, and understand why its pursuit is facilitated by their presence. To distinguish these two aspects of the problem, of course, does not imply that they are resolved by distinct processes within an opportunistic planner: Recognizing the features that constitute an opportunity may be just as dependent on recognizing the goal for which they are an opportunity as the other way around. But however it is accomplished, the recognition of opportunities must contend with the fundamental problem that they often seem to arise when they are not expected. Thus, recognizing potential opportunities entails noticing that certain features of a situation are relevant to the pursuit of goals other than those which govern the agent's current behavior in that situation.

Once those features of a situation which constitute an opportunity have been recognized, and the goal for which they are an opportunity have somehow been "brought to attention," an agent must devote some effort to determining how good the opportunity is, whether, and if so, how, it should be pursued, and so on. However, although the reasoning involved in such decisions is complex and by no means well understood, the functional rationale for such processing within an opportunistic planner depends first of all on being able to recognize opportunities. Moreover, although an opportunistic approach will undoubtedly have an impact on how and when such processing is carried out, in a more general form these are issues which must be addressed in any theory of planning. The central and unique problem which must be addressed by a theory of opportunistic behavior, therefore, is how to detect opportunities, and

"activate" the goals to which they pertain. That is what this chapter is about.

Several difficult issues arise in attempting to address the problem of recognizing opportunities. First, there is simply the problem of recognizing the presence of those features which constitute the opportunity. How much and what kind of processing is required in order to recognize the presence of such features, and why is that effort expended in any given case? If the features are simple enough or important enough, it may be the case that the agent can carry out the processing necessary for their recognition at all times, regardless of context. In general, however, this will not be possible: If a relevant feature is sufficiently complex, the processing necessary to recognize it may be computationally too expensive to carry out unless there is reason to believe that it would be useful to do so. In other words, in most cases there must be some specific reason why such processing would be undertaken. There are two possibilities: Either the processing necessary in order to recognize a feature has been performed in the service of some *other* goal which is currently being pursued by the agent, or the goal for which the feature constitutes an opportunity *itself* plays some role in the recognition of that feature. In any case, the greater the amount of effort required to recognize the presence of some feature which constitutes an opportunity, the more pressing becomes the question of why that effort has been expended. Moreover, even if the recognition of a feature seems relatively unproblematic, it does not follow that it will be trivial to realize that the feature in question actually does constitute an opportunity, and to activate the relevant goal. Again, the greater the effort required to recognize that some feature constitutes an opportunity, the more pressing the question of why that effort has been expended.

To some extent, the distinction drawn in apportioning the effort necessary to establish that an opportunity exists between "feature detection" and "goal activation" is somewhat artificial. Regardless of how the pie is cut, however, the more effort that is necessary to detect a given opportunity, the more important -- and more difficult -- it is to justify why that effort was made. There is, in other words, a great deal of pressure to construct models for the recognition of opportunities which minimize the effort required. On the other hand, it seems likely that the subtlety of the opportunities which can be detected will depend on the amount of effort devoted to that task. In view of this trade-off, in this chapter I will present several different architectures for recognizing opportunities, varying in the amount and kind of effort devoted to the problem, and also varying, as a result, in the subtlety of the opportunities they are capable of recognizing. Thus, to the extent that the reader believes that *people* are capable of noticing opportunities of a given subtlety, that ability constitutes evidence for the necessity and feasibility of an architecture capable of a similar level of performance.

2. The "mental notes" model

Opportunistic behavior depends on the ability to activate goals under circumstances which facilitate their pursuit. Thus, goals must be linked in memory to descriptions of situations in which such opportunities might arise. They must, in other words, be indexed in memory in terms of features which indicate opportunities for their achievement, so that they can be activated upon the detection of such features. Thus, the question of *where* to store a goal in memory depends on *when* it should be activated. When will an opportunity to pursue an unsatisfied goal tend to arise? The most obvious answer is when the conditions that prevented its immediate satisfaction -- usually, the absence of some necessary preconditions -- no longer hold. It follows, then, that whenever a goal is formed, if it cannot be immediately satisfied, it should be indexed in terms of the unmet preconditions that prevented its satisfaction.

Suppose, for example, that you needed some item -- e.g., a technical book -- that you didn't possess. In order to possess the item, you might need to go to a certain store, say the Yale Co-op, in order to buy it. You might just go to the store immediately and purchase the item. On the other hand, that might be impossible at the moment or else not worth the trouble. Where then should the unsatisfied goal "buying a technical book at the Yale Co-op" be stored in memory? On this approach, it would be indexed in terms of the missing precondition "agent is at the location of the Yale Co-op." In other words, you would just make a "mental note" that the next time you passed by the Co-op, you should go in and buy the item.

Opportunities to pursue a pending goal may also exist whenever other, similar goals are active, since the conditions under which they are likely to be pursued are, on account of their similarity, likely to facilitate the pending goal as well. Thus, an unsatisfied goal should be indexed with other, similar goals in the planner's memory, so that it can be easily brought to mind when those other goals are contemplated or pursued. For example, the goal of buying a technical book at the Co-op would, presumably, be represented in memory in terms of -- and could therefore be indexed under -- more general goal/plan structures, such as "shopping at the Co-op," or "buying a hard-to-find item," which in turn would be represented in terms of a yet more general "shopping" goal/plan structure. That is, the unsatisfied goal of buying a technical book at the Yale Co-op should be stored with other, similar goals in a goal/plan hierarchy in the planner's memory. Whenever the planner then happened to form the goal of buying something *else* at the Co-op, or happened to be shopping at the Co-op, the goal/plan structures that deal with this situation would presumably be retrieved or activated to guide his planning or behavior. And since the unsatisfied goal would be stored there as well, it too could be

activated.

One of the major strong points of the "mental notes" model is its simplicity, in that it does not attribute very much processing power to the pending goals themselves. Rather, the whole idea of this model is that goals are brought to the attention of some *other* process, a central planner of some kind concerned with determining priorities among active goals and constructing plans to pursue them. Even so, of course, some processing must be devoted to the detection of opportunities. First, there is the processing that is required simply in order to recognize the feature or cluster of features which constitute the opportunity, and in terms of which the goal is therefore indexed. On this approach, however, the presence of the goal itself plays no role in that processing: It must be a natural consequence of processing that is being carried out for some other purpose. For example, recognition of the feature might be an inevitable consequence of the processes used to perceive and understand in any situation. Or, more generally, it might be that recognition is dependent on the coincidental influence of some *other* goal which is also concerned with that feature, but for some other reason. In those cases when a goal is activated because another, similar goal has been activated, almost all of the relevant features of the situation -- e.g., buying something at the Co-op -- would be recognized, and the right place in memory accessed, all in the service of that other, similar goal. Whether due to context-independent processing or the influence of some other goal, however, in the mental notes model the recognition of a feature which constitutes an opportunity is entirely "bottom-up" with respect to the goal for which it constitutes an opportunity

Although the mental notes model does not entail devoting any special effort to the detection of features that might constitute opportunities, activating the goals for which they constitute opportunities does require a certain amount of special processing. Even if an instance of the feature (or cluster of features) in terms of which a goal is indexed has been detected, it does not automatically follow that the goal will be activated or retrieved. There are likely to be many other concepts, features, or structures indexed in terms of or somehow associated with the given feature. However, an exhaustive investigation of all of these structures seems unrealistic. Moreover, "spreading activation" or "marker passing" -- while perhaps relevant to the general problem of opportunistic planning -- are not relevant here. Even supposing that all concepts associated with a given feature are somehow "activated" or "marked," they must still be investigated by some other process to determine whether or not *there is a pending goal among them.*

Thus, some sort of special status must be bestowed upon the pending goals which are

indexed in terms of a given feature or structure, so that if an instance of that feature is detected, the pending goals in particular will be brought to the attention of the planner as a whole. There are several ways in which this could be accomplished. For example, a special "pending goal" link might be used to connect goals to those features which indicate an opportunity for their pursuit. Whenever an instance of a feature were recognized, all of the "pending goal" links attached to that feature, if any, could be investigated, and the pending goals to which they point could then be activated and brought to the attention of some central planning process (unless perhaps there were a crisis which demanded all of the planner's attention). Alternatively, a pending goal might be able to call the central planner's attention to itself whenever the feature in terms of which it is indexed were recognized, e.g., by placing itself on a queue of active goals.

These two methods are, essentially, equivalent. The first is a bit more "serial" in flavor -- it is based on the notion of "polling" -- while the second seems a bit more "parallel" -- it is based on the notions of "demon" and "interrupt." But in either case, pending goals are singled out as special entities to which some extra processing is devoted simply by virtue of the fact that they are pending goals. Thus, even in its simplest form, opportunistic behavior requires attributing a certain amount of processing power to goals in order to recognize opportunities for their pursuit. Sufficient attention must always be devoted to pending goals to enable their activation upon the recognition of the features in terms of which they have been indexed.

3. Mental notes: Elaborate and index

The mental notes model which was presented in the last section is rather simple, probably too simple to account for the opportunistic abilities of human beings without substantial modification. Consider, in particular, the case in which some opportunity to pursue a goal presents itself *other* than the one that was originally planned for. For instance, to continue with our example of buying a technical book at the Co-op, consider all of the other ways in which the goal of gaining possession of the book might be accomplished: You might happen to pass *another* store that might have the book that you need, in which case you could buy it there. Or, someone might tell you that they plan to go to the Co-op, in which case you could ask them to pick up the book for you. Or, you might run into a colleague who might have a copy of the book, in which case you could borrow it from him. Or, indeed, you might encounter someone who can tell you what you needed to know from the book, and by thus satisfying the higher goal which gave rise to the intention to buy it in the first place, render its purchase unnecessary.

The import of such examples is clear: If the conditions which constitute an opportunity

to pursue some goal have been characterized too specifically -- if, in other words, the goal has been indexed in terms of overly specific features -- then there will be many cases in which it will not be aroused even though an opportunity for its satisfaction is present. On the other hand, characterizing the opportunity to pursue a goal too abstractly -- that is, indexing the goal in terms of highly abstract features -- makes it much more difficult to retrieve the goal under *any* circumstances: Such abstract features will tend to be more difficult and expensive to recognize, and so it is less likely that the processing necessary to do so will be carried in any given situation. For example, we can presume that if an agent is in close proximity to the Yale Co-op, and it knows what the Yale Co-op is, then the Yale Co-op will be recognized -- i.e., the feature "Yale Co-op" will be recorded as present, or "activated." But if the feature in terms of which the goal is indexed is more abstract -- e.g., something like "source of technical books" -- then we cannot assume that the agent will automatically generate such an abstract description for any given bookstore, library, or colleague which he happens to pass by. Unfortunately, if he does not, then the goal will not be aroused.

The dilemma then, is this: If a goal is indexed too specifically, opportunities for its achievement may be missed, since the specific feature in terms of which it is indexed will not, in all probability, be present in many cases in which an opportunity nevertheless exists. On the other hand, under the mental notes model, a pending goal does not play any role in the recognition of the feature which signifies an opportunity for its achievement. Thus, such recognition is entirely dependent on processing that is carried out either automatically or coincidentally in the service of some other goal. But the more processing that is required, the less likely it is that such processing will be carried out either automatically coincidentally. Thus, if a pending goal is indexed too abstractly, the feature in terms of which it is indexed will, in all likelihood, not be recognized even in many situations in which it happens to be present. Once again, therefore, many opportunities are likely to be overlooked.

One solution to this dilemma within the framework of the mental notes model is to spend some effort, when a goal is formed, to determine a number of situations in which it might be easily satisfied -- for example, by constructing several incomplete plans for the goal in order to identify the relevant preconditions -- and then index the goal in terms of all of the features that might arise in such situations (this idea is originally due to Dehn, in preparation). Thus, rather than just thinking of one plan that might enable you to achieve some goal, you would think of as many as you could, and then go about indexing the goal in terms of all the preconditions for all of those plans -- or at least the problematic ones. For example, in addition to the Co-op, you might also index the goal to get some technical book in terms of features representing certain other bookstores, as well as the engineering library. Such additional work might also

be done incrementally whenever the goal were active for any reason. By indexing a goal in terms of several different opportunities in this way, the probability of actually recognizing an opportunity for a given goal is improved, without depending on the derivation of complex, abstract descriptions of situations in which the agent finds itself. On the other hand, even with this elaboration the mental notes model has certain inherent limitations: It does not enable an agent to recognize opportunities other than those which it has anticipated -- and which, *a fortiori*, it is able to anticipate -- and for which it has prepared by appropriately picking the features in terms of which to index its goals. In particular, therefore, it does not enable an agent to recognize *novel* opportunities, which, by their very nature, cannot be anticipated. I will return to this problem later in the chapter.

4. Structured features and the two-tier model

On any account of opportunistic behavior based on the mental notes model, a pending goal cannot be indexed in terms of a particular *instance* of some feature which constitutes an opportunity, since the agent cannot know in advance the particular situation in which the opportunity will arise. Rather, it must be indexed in terms of the *type* of the feature which indicates an opportunity for its pursuit. If any instance of that type were detected, the goal would then be brought to the attention of the planner as a whole. However, in the mental notes model, no particular effort is expended to recognize an instance of such a feature, since the pending goal itself plays no role in the recognition of the feature or features in terms of which it is indexed. Under such conditions, the process of recognizing an instance of the feature -- in other words, of establishing that it is an instance of the given type of feature -- cannot depend on the application of sophisticated and expensive processing, since there is no guarantee that such processing will in fact be carried out in any given case. Thus, the feasibility of the mental notes model depends on being able to recognize, rather immediately, an instance of the feature type in terms of which a goal is indexed.

In many cases this is not terribly difficult, because the description of the situation may include an explicit representation of the type of the feature -- that is, an instance of that feature type is represented by directly invoking the *name* of the type itself. For example, if the opportunity depends on the presence of an instance of the feature type "Red," and if the situation is described as containing an instance of that feature -- that is, if red is present -- then that description will entail the use of precisely the symbol which stands for the feature type "Red." In other words, once a description of some situation has been assembled, the presence of instances of the feature types which are explicitly and directly employed in the representation of that description can be trivially determined by inspection.

What if this is not the case, however? An instance of some feature type may implicitly exist in the representation of some situation even though the symbol associated with the type itself is not explicitly present in that representation. In such cases, determining that the situation includes an instance of the feature type -- and explictly representing that fact -- will require further inference, even though no additional information about the situation needs to be included in the description in order to make that determination. The importance of this problem for opportunistic planning is that such an explicit representation of a feature type seems necessary if that feature is to be employed as an index in memory, and thus permit the activation of pending goals for which the feature constitutes an opportunity (although see Hopfield, 1982, and Ackley, Hinton, and Sejnowski, 1985, for some recent attempts to construct memory models that do not depend on the use of explicit names in indexing). That is, it is not enough in the mental notes model that some feature be implicitly present in the description of the situation -- to be used as an index, its presence must be *explicitly* determined and *explicitly* represented.

It may seem, at first, difficult to understand how the recognition of an instance of some feature in some situation could be problematic when all of the information necessary to determine its presence is already available in the representation, or even more, when all of the components that make up the feature in question are already themselves explicitly represented. But this is exactly the case when the feature in question is a complex *structure* of other features. Let's consider once again our example in which the agent's goal is to buy a technical book at the Yale Co-op. The actual feature of the world which constitutes an opportunity to satisfy this goal is not simply "Yale Co-op," but rather "Agent is near the Yale Co-op." This feature is a complex combination of other features -- e.g., "agent," "near," and "Yale Co-op" -- and it is not clear that such a combination of features would automatically be constructed even if the Yale Co-op were recognized. That is, the representation of the situation might not even include all of the information necessary in order to recognize such a complex feature. If it did not, then the agent would not be able to recognize the opportunity. In this particular case, of course, it is likely that such a description would in fact be constructed simply for navigational purposes, because an agent trying to get around in the world would naturally be constructing such descriptions about its location in order to plan and monitor its route.

However, even if all of the information necessary is present, and such a description is constructed, it does not automatically follow that the feature types to which it corresponds, and in terms of which the goal to buy the book is indexed, will in fact be called to mind. Even leaving aside the recognition problems posed by such abstract features as "source of technical books," it usually takes a fairly sophisticated form of pattern-matching -- indeed, it requires an

inference -- simply to recognize that a given feature applies to a given situation. Such an effort is necessary as soon as the features in question are not explicitly named in the description of the situation, but rather comprise some structured combination of such features. A planner might describe a situation in terms of the complex representation "Agent is near the Yale Co-op," and *still* not recognize that this description exactly matches a complex feature in its memory. For unlike a feature like "Red," or even "Yale Co-op," this instance of the feature need not use the exact same *symbol* as the corresponding type indexed in memory. It is true that both descriptions use the same component symbols -- "Agent," "near," and "Yale Co-op" -- but this alone does not guarantee that the matching description in memory will be recognized and retrieved. Moreover, the more abstract a feature is, the more likely it is that even its components are not explicitly represented in the input, even when all of the information necessary to establish their presence, and its own, is in fact explicitly represented.

To put this another way, any process that is capable of determining the presence of a complex, structured feature, must be sensitive not only to the presence or absence of its component features, but to the structural relations among those features. For example, merely from the presence of the component features "agent," "near," and "Yale Co-op," one cannot infer that the "agent is near the Yale Co-op." It might be, rather, that the agent is near his television set, and watching a story about the Yale Co-op on the evening news. Even the presence of the features "buy" and "Yale Co-op" are not themselves sufficient to indicate that the goal/plan structure "buying something at the Yale Co-op" is applicable: One might be reading a story about the Yale Co-op buying a new building. Complex or abstract features are not merely unstructured bundles of more primitive, component features -- rather, they are *structured* sets of such features. The component features out of which they are constructed must bear certain relationships to each other. In order to keep such relationships straight, an inferential memory system must be able to manipulate variables and variable bindings -- which is to say, it must have the symbol-manipulation abilities of such programming languages as Lisp or Prolog. The process of detecting complex or abstract features such as "agent is near the Yale Co-op," or "place where technical books are available," depends on having this capability and employing it in the recognition of such features.

But the need to employ such sophisticated pattern-matching capabilities immediately poses a severe problem for any version of the mental notes model: Why should such processing be performed? Unless there is some active goal which is specifically concerned with a given structured feature, it seems hard to justify expending such an effort to detect its presence and explicitly represent it. Indeed, it is to avoid exactly this sort of problem that model was elaborated to include the multiple indexing of goals under features which are not

overly difficult to recognize.

To some extent, this problem can be sidestepped if the feature in question is of interest to another, currently active goal, and this is a reasonable expectation if that goal is itself very similar to the goal for which the feature might constitute an opportunity. Thus, there is no particular difficulty in recognizing that the description "agent is at the Yale Co-op" matches the precondition for buying a technical book at the Yale Co-op if one is going to the Yale Co-op to buy something else, because that feature is relevant as a precondition for the active goal. In general, however, it cannot be expected that the feature which constitutes an opportunity for one goal will also be of interest to another goal that happens to be active. For example, if the agent simply passes the Co-op on the way to some other destination, he may fail to recognize that the precondition for buying the book he wants has been satisfied, and that an opportunity to pursue that pending goal therefore exists.

Probably the simplest solution to this problem is to index goals in terms of single, specific features such as "Yale Co-op," instances of which are likely to be easily detected in a description of the situation, rather than in terms of complex, structured features such as "agent is near the Yale Co-op." However, the use of simpler features in this way will, inevitably, result in the goal being called to mind in many circumstances which do not actually constitute an opportunity for its pursuit. For example, suppose one is reading about the Yale Co-op's financial situation in the local newspaper. The feature "Yale Co-op" will be present in this situation. Should the goal of buying a technical book at the Co-op then be called to mind? Suppose one passes the Co-op late at night, or on Sunday, when it is closed. Should pending goals be called to mind under such circumstances? Are they, in fact, in human beings?

If goals are indexed in terms of such relatively unstructured features -- which are more easily detected -- then, as the examples above illustrate, we can assume that there will be many false alarms. Thus, the feasibility of this approach depends on how difficult it is for the planner as a whole to determine whether or not a goal which, superficially, *looks* relevant in the current situation, is in fact relevant. That is, this approach leads to a partition of the detection problem into two phases. One is a simple noticing process which tends to overdo it -- that is, it aims for completeness and thus gives many false alarms. The other is a more cautious and expensive inferential phase, which has the task of determining whether or not a *bona fide* opportunity exists. For example, whenever the feature "Yale Co-op" were present in some perceptual input -- or even in some plan or thought of the agent's -- the goal of buying the technical book would be aroused. Inferential processing would then be employed to determine whether or not the actual precondition which involves this feature -- namely,

proximity to the Yale Co-op -- in fact held, and therefore whether or not an opportunity actually existed. If so, then the agent would recognize an opportunity, and decide whether or not it should be pursued. If not, then the goal would be dismissed. (A goal might also be dismissed as not sufficiently important given other currently active goals without even bothering to determine whether an opportunity actually existed.) In effect, under this elaboration of the mental notes model, the problem of noticing an opportunity is approached using a generate and test architecture. Goals are activated by the presence of instances of easily detectable feature types in terms of which they are indexed. They are then brought to the attention of the planner as a whole, which in turn determines whether or not an opportunity actually exists using more sophisticated inferential processing.

In sum, the mental notes model attempts to minimize the attribution of inferential power to the pending goals themselves, avoiding the problems and costs inherent in the recognition of complex features by stipulating that such goals are indexed in terms of relatively simple features that do not require the heavy use of inference in order to be detected and explicitly represented. However, in order to avoid missing many opportunities, the use of relatively simple features in this way will result in a tendency to overestimate the presence of opportunities. Thus, once a goal is aroused on account of the presence of some simple feature that might indicate an opportunity, more sophisticated *inferential* processing must be employed to determine whether or not the complex, structured feature that would actually constitute an opportunity is in fact present. This is a *two-tier* model of opportunity recognition.

5. Goal arousal and inferential processing

On the two-tier model sketched out in the last section, a pending goal can be in one of two possible states of arousal, distinguished by the computational resources devoted to that goal. If a goal is in the low state of arousal, no particular effort is being made to detect the features that might constitute an opportunity for that goal. However, sufficient attention is always devoted to a goal in this state so that, should one of those features be encountered, it will be recognized, the goal will be elevated to a high state of arousal. This high state of arousal is distinguished by the fact that if a pending goal is in such a state, specific efforts are undertaken towards the recognition of complex features that constitute potential opportunities for its pursuit. In other words, inferential resources are devoted to a high-arousal goal just as they are to a goal under active planning or pursuit. Indeed, the attempt to determine whether or not a genuine opportunity exists is itself a planning task. The claim that a state of high arousal is in fact equivalent to a decision to [illegible]

The distinction between a low and high state of arousal immediately raises the question of whether or not there are other reasons why a goal should be elevated to the second level, that is, endowed with the inferential resources necessary to recognize complex features of specific relevance to it, beyond (a) the fact that it is currently being pursued or part of the detection of a simple feature which may indicate an opportunity for its pursuit. The answer, I will argue below, seems to be yes. However, if a goal has been elevated to a high state of arousal in order to determine whether or not a genuine opportunity exists, the purpose of devoting such inferential effort is clear. If a goal has been elevated to the second level for some other reason, the situation is not so straightforward.

What might such a reason be? The most [illegible] a state of high arousal because it bears some [illegible] that state. For example, it might be [illegible] the arousal of the subgoal [illegible] planning for or pursuing [illegible] planning and plan execution. [illegible] accommodate the needs of [illegible] that resembles [illegible] goals are [illegible] be used [illegible]

accurately, to pursue the main goal without having to suffer one or more of its distasteful aspects -- then it is important for the planner to recognize such an opportunity, and further, to determine whether or not the conditions that led to that opportunity can be replicated in the future. For example, you might throw something in the trash that turns out, unexpectedly, to neutralize bad odors, and then perhaps realize that you can use this substance in the future whenever the trash smells bad.

One of the most important aspects of the story sketched out above is that the opportunity which is recognized to pursue the main goal while avoiding the unpleasant consequences is *novel* -- the planner did not, and could not, prepare for it ahead of time. No effort was spent indexing the frustrated goal in terms of relevant preconditions, because the planner was not originally aware of any feasible plans by which it could be pursued. Rather, the recognition of the opportunity in this case depends on two factors: The failure of the expectation that the garbage will smell bad -- or to put this another way, of the expectation that the goal of avoiding bad odors will be frustrated -- and the recognition that this failure resolves the conflict. In other words, the opportunity arises because the situation presents us with an unexpected resolution of the goal conflict between taking out the trash and avoiding bad odors. In a sense, the goal for which an opportunity exists is the goal to resolve this goal conflict. The recognition that such a highly abstract goal has been or may be fortuitously satisfied, however, depends on the computation of a highly abstract and structured representation of the situation, one that clearly requires a heavy commitment of inferential processing.

[illegible] point to be made here is that pending goals may be in different states of arousal [illegible] differences are in part defined by the kind of effort that is made to notice [illegible] In particular, the allocation of inferential resources to a [illegible] recognition of more abstract and highly structured features which [illegible] permit the recognition of more novel [illegible] and reasons to suppose that [illegible]

avoided, while an opportunity may be fleeting and similarly require split-second decisions to be seized. Thus, given that many features of a situation may be relevant to the pursuit of many goals, it is clearly necessary to devote effort to considering more important goals first. In the simplest case, for example, more important goals might have a "louder voice" in calling themselves to the attention of the planner as a whole.

Once the factor of importance is taken into account, the possibility is also raised that some goals are so important that extra effort should always be devoted to recognizing opportunities for their pursuit or threats to their satisfaction. That is, there is good reason to believe that some goals may be so important that they should *always* be aroused and active. For example, in an environment where lack of food is a constant problem, a person will probably notice and attempt to pursue almost every opportunity to get food, even if he is not hungry at the time the opportunity arises.

In sum, taking the mental notes approach, we start by assuming that goals are endowed with the minimum capacity necessary to call attention to themselves when an opportunity seems to exist. However, this leads to a model in which there will be many false alarms, which in turn motivates the application of more sophisticated inferential processing to determine whether an opportunity actually exists. Once we see that such inferential capacity must be allocated to goals which have been aroused because they are potentially relevant, the ways is open to allocate such capacity for other reasons as well: Either because of their relation to other highly aroused goals, or simply because they are deemed very important. This marks a step away from the mental notes model -- in which pending goals are conceived of as "active" only in the sense that they can call attention to themselves if the feature in terms of which they are indexed is clearly present -- towards a model in which pending goals may sometimes or always be active in the much stronger sense that they are allocated inferential capacity dedicated to the task of noticing opportunities for their pursuit.

6. Inference and novel opportunities

As we have seen, the mental notes model has certain limitations as a model of [illegible]. These stem primarily from the fact that in order to notice an [illegible] between the [illegible] opportunities. The best [illegible] opportunities [illegible]

-- simply cannot be recognized.

For example, suppose that you need to know about something and decide to get hold of some particular book to find out about it. You might form the plan of buying the book at the Co-op. But, will you be able to recognize when another opportunity presents itself? What about another bookstore? What about a library? What about a colleague's private collection of books? What about a new plan altogether -- such as finding out what you need from a knowledgeable colleague directly? In these cases, it seems clear that the goal cannot be indexed directly in terms of the feature that constitutes the opportunity -- or else, if it is, that feature is highly abstract and far removed from perceptual inputs.

Even though a novel opportunity for a goal cannot, by definition, be directly linked to the feature that constitutes the opportunity, there must of course be some indirect link between them, in the sense that there is a series of inference rules which connect the feature to the goal in some way. Otherwise, the agent would never be able to recognize the opportunity at all. This is just another way of saying that the agent must have the requisite causal knowledge to understand how the feature facilitates the achievement of the goal if he is to recognize the opportunity.

Once it is granted that some kind of path exists in memory between the feature and the goal, however indirect, it is tempting to rely on something like "spreading activation," or "marker passing," or some other sort of automatic memory search mechanism, to somehow trigger the feature in terms of which the goal was originally indexed. Thus, for example, if the feature were quite abstract -- "place where technical books can be found" or "person who knows about subject X" -- then we might simply assume that if we pass a place or that person that had such properties, those features would automatically be recognized, or at least aroused in some way. Or, if the feature were quite specific -- e.g., "Yale Co-op" -- then upon passing another large bookstore, the features that it shares in common with the Yale Co-op would "remind" the agent of the Co-op, and hence activate the goal.

There are several problems with such an approach, however. First, it is always possible to sketch out a story for the recognition of some opportunity using marker passing or spreading activation, simply by postulating that the appropriate links exist, and that too many of the "inappropriate" links do not -- so that the goal in question will be activated without activating so many other goals that the system becomes hopelessly bogged down. How well it could be expected to work in a highly interconnected memory with many goals is, however, open to question.

Second, these methods still depend on pre-existing connections between structured features and their components, so that activation or marking of a component or components leads to activation or marking of the structure in which they play a role. What I now want to show is that there are features instances of which *cannot* be recognized using marker passing or spreading activation over pre-existing connections, because they are abstract features which have no pre-existing connections with their components in any given instance -- or at least, with any of the components which are likely to be themselves explicitly present in the representation. Inferential processing must therefore be devoted to the task of recognizing such features if at any time the agent wishes to determine their presence. Thus, an opportunistic planner which can recognize opportunities on the basis of such features must devote inferential effort to detect the opportunity, either in the service of the goal for which the opportunity exists, or in the service of some other goal.

Let us see how to construct such features. They must be structured patterns of more primitive, component features, in which no single component feature can be expected to be directly linked with the larger feature of which it is a component. Consider phenomena such as punning or *doubles entendres* -- the latter are simpler so we will stick to them. A *double entendre* is an utterance with two meanings, one usually not quite socially acceptable, both meanings of which are intended and significant in the context. A double entendre is usually considered humorous, and the person who utters it witty.

First of all, it should be clear that *doubles entendres* are a good example of the sort of capability that opportunistic planning is intended to account for. A speaker cannot, and clearly does not, always plan out such utterances in advance. Rather, what seems to be the case is that the speaker starts to say something, and then realizes that, if he completes the utterance appropriately, he will be able to convey two distinct but salient meanings in a humorous fashion. The precondition for a *double entendre*, then, is that some word or phrase in the utterance has two meanings, both appropriate in the context, but appropriate in somewhat different ways. The speaker starts to plan out the utterance of some thought and realizes that his choice of words affords such an opportunity to utter a double entendre.

Thus, most *double entendres* fall into the category of [illegible] opportunity for a *double entendre* is discovered, there [illegible] that constitutes the opportunity and the goal to be [illegible] see why this must be so, consider what the [illegible] which signifies an opportunity to utter one [illegible]

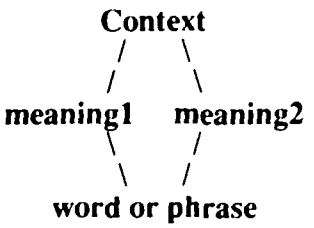


Figure 8-1: Precondition for a *double entendre*

Now, the point here is that this precondition does not depend on the presence of any particular feature -- not the presence of the particular word or phrase, or the particular meanings, or the particular context. Rather, its essence lies in the way that all of these elements are related to one another. Thus, all possible instances of this pattern cannot possibly be connected to the general type of the concept. That would entail wiring up all possible combinations of words and phrases -- an impossible task in and of itself, given that all phrases cannot possibly be explicitly represented in memory -- their meanings, and contexts in which they might be relevant, such that if all the right primitive features were present -- this word, those meanings, that context -- then, and only then, would a double entendre be present. This is simply absurd on the face of it. Thus, the recognition of a *double entendre*, or of its precondition, is exactly the sort of problem that "spreading activation", "marker passing" -- and, at least as they are currently formulated, connectionist models -- are not capable of dealing with. The recognition of such a feature requires inference -- it requires taking the general pattern, binding variables, and checking to see that the constraints hold in the given situation.

Thus, the feature that indicates an opportunity to produce a *double entendre* cannot be expected to show up automatically in the description of the situation in which a speaker finds itself. In order to recognize the above pattern, the agent must have a mechanism dedicated to finding it among the words he plans to speak. It can, however, be argued that its recognition is a side effect of a speaker's goal to be clear and unambiguous. If a speaker has such a goal then it is possible that he monitors his planned speech in order to check for ambiguities which are difficult to resolve -- for example, because both meanings are sensible in the context -- and edit any such speech in order to be clearer. The precondition for a *double entendre* is very similar to such a situation, differing only in that the ambiguity is humorous rather than merely confusing. If, on the other hand, the above scenario does not hold -- if, that is, the speaker is not routinely checking his speech for unresolvable ambiguities in order to avoid confusion -- then the only conceivable motivation for recognizing opportunities for *double entendre* is that he is interested in being witty. In other words, it is possible that the goal of being humorous must be activated explicitly in order for the opportunities for its

achievement.

Ultimately, a given goal may not be able to rely on the automatic recognition of features that are relevant to its pursuit, or on their recognition in the service of some other goal. In such cases, if opportunities to pursue the goal are to be recognized, then the goal itself must play a role in their discovery. The mental notes model is bottom-up with respect to noticing opportunities for goals, and this may not work in general: It may be necessary that a goal play a top-down role in the identification of features relevant to its pursuit, even though it is not currently under active pursuit itself. This, in turn, would entail attributing far more inferential power to the goals themselves. The allocation of such resources need not be unvarying over time, however. Some goals are more important than others, and the relative importance of a goal may change depending on context. For example, the goal of being humorous -- and being deemed so by others -- is likely to be more important at a party than, say, in a private discussion of some technical issue. It seems entirely reasonable, then, that under such circumstances a speaker may be more sensitive to the presence of an opportunity to utter a *double entendre*.

Let's consider another example of this sort. When an AI researcher is reminded of some *situation by some other situation, and then realizes that the reminding* belongs to some interesting class -- e.g., the reminding has no low level features in common with the input -- it cannot be that the reminding was previously classified in that way. After all, the relation between the reminding and the input is a dynamic property of the mind -- it didn't even exist before the reminding occurred. Thus, the researcher must have had the goal of looking for such remindings actively. The concept 'reminding with no low level features in common' cannot be connected to all such remindings, for that simply cannot, in principle, be done ahead of time. Rather, it must be done 'on the fly.' Even explicitly noticing that you have been reminded at all -- as opposed to simply being reminded -- cannot be taken for granted. The recognition by an agent of features of his own dynamic state of mind is a significant act of perception.

This section has discussed the effort required to recognize opportunities for rehearsal when [illegible] as discussed in the previous chapter. In the simplest cases, an opportunity [illegible] However, [illegible] by its memory that are associated with the goal. That is, [illegible] Finally, [illegible]

memory. Thus, even when those facts are recalled, it will require inference to explicitly notice the contradiction. Noticing more subtle incongruities -- e.g., that a related fact does not contradict a claim, but does in fact reduce its force as evidence for some other claim -- will require even more work.

The point here is this: Features in the world that are relevant to the pursuit of goals are in fact highly structured, often highly abstract, sets of features. They cannot be represented simply as "feature vectors," that is, as unstructured sets of features, because it is often the relationships among the component features involved that are most crucial. Now, it might be argued that such relationships can themselves can be counted as features, and, therefore, added to the feature vector describing the situation at hand. This is true, but it misses the point. It is not merely the *presence* of these relationships that matters: It is the fact that they relate certain other features of the situation in particular. Thus, simply saying that a situation involves not only the feature "agent" and "Yale Co-op," but also the relationship "near" as a feature as well, is not the same as saying that this relationship holds between the first two features. This fact too must be represented to get a causally meaningful representation of the situation. Thus, the relationships among features must be explicitly represented, and the scope of the relationships -- which features in particular they apply to -- must be explicitly represented as well. This, in general, requires the ability to manipulate general symbolic structures: Unstructured "feature vectors" will not suffice.

What does this say about opportunistic planning? Since the features which indicate opportunities cannot, in general, be represented as single features or simple bundles of such features, the processes which notice the presence of such features must be able to construct and manipulate general symbolic expressions. This is not cheap: It cannot be done by such simple processes as spreading activation or marker passing. Rather, it requires inferential processing to construct such complex, structured features. The above examples, in which spreading activation or marker passing cannot possibly work because the links which are required to spread activation or pass markers do not even exist, makes this point quite clear.

Thus, structured features must, in a sense, be explicitly looked for: No cheap, "automatic" process will deliver them to the understander. Thus, we have three alternatives as to how the recognition of such a feature might be accomplished. First, it might be recognized at all times, regardless of context. Second, it might be recognized in the service of some goal under current pursuit. Or third, it might be recognized in the service of the goal for which it constitutes an opportunity. The first alternative is highly implausible, since the need to employ inference to recognize such features means that they are expensive to check for. Thus, in order

to recognize opportunities that depend on complex and abstract features, such features must either be of interest to some goal under current pursuit, or else the goal for which they constitute an opportunity must play a role in their detection.

7. Conclusion

In this chapter, we have begun to explore some of the architectural requirements for opportunistic planning. Starting with a simple mental notes model -- in which goals are directly indexed in terms of the features that constitute opportunities for their pursuit -- we saw that goals must be active agents to the extent that they can call attention to themselves whenever the features in terms of which they are indexed are recognized. I then argued that, since the features that actually constitute opportunities are often highly structured or abstract, it might not be practical to assume that they would be recognized bottom-up. This led to the proposal that goals should be indexed in terms of simple features that are relatively easy to recognize. However, I argued that this in turn would lead to a situation in which there would be false alarms noticing opportunities. The solution was to postulate that, once a goal was aroused because a simple feature that potentially indicated an opportunity had been recognized, inferential resources would be allocated to checking whether or not an opportunity truly existed.

Finally, we saw that certain kinds of features cannot even be recognized as potentially present without the application of inference. If the opportunity to pursue some goal depends on such features, there is no relatively simple feature in terms of which to index the goal so that it can be opportunistically aroused. In such cases, if an opportunity is to be detected at all, inferential resources must be allocated to the task. Either those resources must be allocated in the service of some other goal, or else the goal for which they constitute an opportunity must itself play an active role in their recognition. In the next chapter, we will see that this notion has some surprising applications.

CHAPTER 9

GOALS AS ACTIVE MENTAL AGENTS

1. Introduction

Throughout this thesis, we have seen that a commitment to mental determinism -- the idea that our thought processes do not take arbitrary turns -- leads to the need for some way to make decisions in a motivated fashion, in both planning and understanding. Such a position has many historical antecedents. Within psychology, Freud was perhaps the chief proponent of mental determinism, and of the use of intentional explanations in order to avoid arbitrariness in psychological theory. The psychoanalytic explanation of errors is in many ways the most striking example of his commitment to this principle, since here Freud attempted to show that "accidents" in human behavior are anything but accidental. Consider his discussion of slips of the tongue in the following passage (Freud, 1935):

> ...the question [must be] raised why just this particular slip is made and no other: one can consider the nature of the mistake. You will see that so long as this question remains unanswered, and the *effect* of the mistake is not explained, the phenomenon remains a pure accident on the psychological side, even if a physiological explanation has been found for it. When it happens that I make a mistake in a word I could obviously do this in an infinite number of ways, in place of the right word substitute any of a thousand others, or make innumerable distortions of the right word. Now, is there anything which forces upon me in a specific instance just this one special slip, out of all those which are possible, or does that remain accidental and arbitrary, can nothing rational be found in answer to this question? (p. 31)

Freud's answer, in the affirmative, depended crucially on an appeal to the speaker's *goals*.

2. Opportunistic planning and Freudian slips

Freud's study of the psychology of errors (see, e.g., Freud, 1935), including notably slips of the tongue, led him to the conclusion that many such errors are not merely the result

of random malfunctions in mental processing, but rather are meaningful psychological acts. That is, they are *intentional* actions in every sense of the word, reflecting and indeed carrying out the goals, whether conscious or not, of the person who commits them. In particular, Freud argues, such errors stem from attempts to carry out *suppressed* intentions, intentions which have been formed but then in some sense withdrawn because they conflict with other, more powerful intentions.

For example, in the simplest case a person may decide to say something, but then change his mind and decide to say something else instead. Nevertheless, the original intention somehow intrudes itself into his utterance. Freud (1935) discusses the example "Dann aber sind Tatsachen zum Vorschwein gekommen," ("and then certain facts were revealed/disgusting"), in which "Vorschwein" is a conflation of "Vorschein" (revealed) and "Schweinereien" (disgusting). The speaker relates that he had originally intended to say that the facts were disgusting, but controlled himself and decided to say something milder instead. In spite of this decision, however, the suppressed intention apparently exerted an influence on his speech.

Examples of this sort show that goals, once formed, can influence subsequent behavior despite intervening decisions to suppress them. Viewed from an information processing perspective, however, there are two radically different interpretations of this fact, corresponding to two distinct models of how the influence might be exerted. On one account, no further processing of the goal is undertaken after its suppression, and the influence is simply a residue of the processing that took place prior to that suppression. In the above example, for instance, it may simply be that the prior contemplation of the goal to say the precise word, "schweinereien," activated that word in memory, and that this residual activation had an effect on the process of choosing what words to say, thus causing the slip. On this account, although the slip does in some sense *reflect* the suppressed goal, it is not really an attempt to carry out the goal.

However, more complex examples show that this sort of [illegible] general, adequate. Consider Freud's example of the toast [illegible] hiccough to the health of our Chief, in which the word [illegible] substituted for the word "anzustossen" (drink). [illegible] is a manifestation of an [illegible] goal on the part of [illegible] superior, suppressed by the social and [illegible] this case in contrast with the simpler examples above, [illegible] speaker's intention [illegible]

word "hiccough:" There are hundreds of *a priori* more plausible words and phrases that can be used to insult or ridicule someone. Thus, the word "hiccough" can only have been chosen in the course of attempting to retrieve the consciously intended word "drink," to which it bears a close similarity in German. Yet, if we accept Freud's analysis of the example, the word "hiccough" was selected because it achieves the speaker's goal to ridicule his superior. Thus,we are forced to conclude that this goal was still active during the attempt to retrieve the word "drink," despite the fact that it was suppressed *prior* to that attempt.

The mere fact that suppressed goals are able to affect the overt behavior of planners is enough to justify the assertion that they are active. However, the sense of activity implied by examples like the above transcends this ability alone. There is no way that a planner could have reasonably anticipated that the goal of ridiculing or insulting its superior would be satisfied by uttering the word "hiccough." However, if the planner was not looking for an opportunity in particular, then it must have been looking for any opportunity to [illegible]. In this case, recognizing the opportunity involved realizing that the substitution of the word "aufzustossen" (hiccough) for the word "anzustossen" (drink) would, in the context of a toast, result in a ridiculous and insulting utterance. Because the effect of the substitution depends on the context, considerable inference is needed to detect the opportunity. It would, indeed, serve to carry out the goal of insulting the superior. Thus, the planner must have expended considerable cognitive resources in detecting the opportunity. [illegible] the goal's formation to the time that it participated in the [illegible].

[illegible]

needed to explain Freudian slips functionally justifiable, or does it merely reflect an accidental attribute of human psychology?

Fundamental to the above explanation of Freudian slips is the ability to recognize and seize unforeseen opportunities to satisfy goals, an ability that I have already argued seems crucial to intelligent behavior. Moreover, we have also seen that the recognition of such opportunities often entails significant inference. This is particularly true if we consider people's ability to seize novel opportunities. It is easy enough to suppose that some features of situations would point directly to goals that they satisfy. For example, it is arguable that, indexed under the feature "money," we have the goal of possessing money. Thus, it isn't hard to see how the opportunity implicit in seeing some money on the street would be recognized.

On the other hand, suppose a person goes to a hardware store and sees a gadget he did not previously know existed, e.g., a router. People seem perfectly capable, at least sometimes, of constructing the inferential chain necessary to recognize how such a novel opportunity might facilitate the achievement of a goal that they could not, ahead of time, have known that it would facilitate. For example, someone who had the goal of possessing bookshelves would seem perfectly capable of realizing that a router would be useful in building them. This seems plausible even if he had *not intended to build the bookcases, but rather had intended to buy* them. In that case, he probably wouldn't have given much thought to how they might be built. But someone who understands what a router does, he may realize that it can be used to cut channels in the side boards of the bookcase, into which horizontal boards can be fitted as shelves.

While the need for this kind of opportunistic processing provides us with a functional explanation for the ability of a goal to recognize the means for its own accomplishment when they unexpectedly present themselves, it remains to be explained why goals which have for whatever reason been suppressed should be able to overcome their suppression when opportunities of the right sort arise. That is, why should an intentional system lack the means to deny certain goals access to mechanisms for producing real behavior?

Sussman points out that opportunistic processing even offers a functional explanation for seemingly unproductive characteristic of an intentional system. Consider the notion of "a goal to be suppressed." A goal would need to be suppressed *if it* were in conflict with another goal in the system. There are two ways that a goal could be in conflict. It could be because the goals themselves are inherently mutually exclusive, or because a different problem arises in attempting to plan for both of them. That is, it is possible for goals to be found to be in conflict based on the planner's judgment of the

resources and options available under the circumstances in which the goals are [illegible] (See Wilensky, 1983, for an analysis of the considerations involved in [illegible] judgments.) For example, the goal of insulting one's boss is presumably suppressed because it conflicts with more important social and political goals. However, the [illegible] goals is situation-dependent. It is perfectly possible that there may be [illegible] which insulting the boss and achieving one's political ends would be [illegible]

Once a goal conflict is recognized, a planner must decide to suppress one goal and pursue the other based on an assessment of which course of action is most reasonable in light of current or expected future circumstances. However, it is quite possible that the circumstances will be different than originally foreseen. Thus an opportunistic planner must be able to override previous decisions about which of its goals to pursue. Decisions made while formulating the plans currently being pursued should not be immutable.

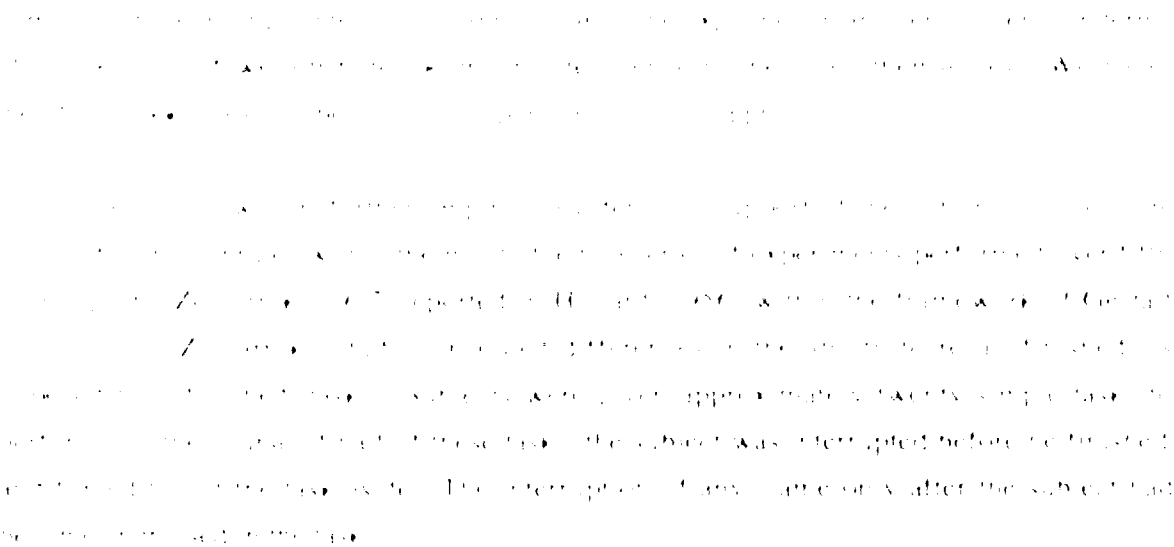
Consider the following example: Suppose a person is out in the forest and is both hungry and thirsty. Given his knowledge about food sources and water sources, and whatever other pragmatic considerations pertain under the circumstances, he decides that these two goals conflict, and that he will suppress the thirst goal while he pursues the aim of satisfying his hunger. While pursuing his plan to obtain food, however, he comes upon a stream which he hadn't previously known about. This is precisely the kind of situation in which we would expect -- or, indeed, demand -- an opportunistic response, regardless of any previous decision to suppress the thirst goal.

The implication here is that the decision to suppress a goal is really just a decision to forgo pursuing that goal for the time being, and that in an opportunistic processor, *no goal is ever really "suppressed."* Viewed in this light, the fact pointed to by Freudian slips, that goals which have putatively been suppressed can still take advantage of opportunities for their own achievement, can not only be understood, but can be seen to be a desirable and possibly necessary aspect of a planner.

What yet remains unexplained, however, is why opportunities would be acted upon even when further reflection by the planner would presumably reaffirm the decision to suppress them, as is undoubtedly the case with Freudian slips. It would seem somewhat counter-productive not to demand that the planner be allowed to reconsider the reasons why a goal was suppressed, in light of the sudden appearance of an opportunity to achieve that goal. One might expect, for example, that despite the opportunity to insult or ridicule one's boss, this opportunity would not be taken, since it would still be impolitic to do so. We might, in fact,

3. The Zeigarnik effect

In the last section, I argued that Freud's intentional explanation of errors entails an architecture for planning and plan execution in which goals are construed as active cognitive agents, endowed with the inferential resources necessary to recognize opportunities for their own pursuit, and I showed that such an architecture is compatible with -- if not in fact necessary for -- real-time opportunistic planning. This view seems to point towards a conception of goals as being inherently distinct from other sorts of mental entities, such as beliefs, rather than simply being data structures on a planner's goal stack, as they are generally conceived to be in AI planning theories. That is, whereas beliefs may well simply be representational structures, goals do not merely represent, they *act*. Such a conception of goals seems attractive in that it does concretely differentiate between two sorts of mental entities which intuitively seem quite different. It does, however, raise the uncomfortable spectre of



Now, this result in and of itself should not really be very surprising. It is clearly useful for planning -- especially opportunistic planning -- that memory facilitate the recall of unsatisfied goals in comparison with satisfied ones, since, presumably, an agent's unsatisfied goals should be of more concern to him in the future than his satisfied goals. Indeed, Hilgard (1956, p. 283) reports that from subsequent experiments, similar in spirit to Zeigarnik's, "it is evident ... that correlated with the memory for the task is a tendency to resume it when the opportunity next arises." Thus, on both functional and empirical grounds, the increased memorability of pending goals is intimately related to the ability to recognize opportunities for their achievement.

The question is, how can a memory accomplish this? Is it a result of clever indexing at the time a goal is formed or frustrated, or is it due to some inherent property of unsatisfied goals? On the indexing account, the difference in the memorability of satisfied goals and unsatisfied goals is simply due to the fact that unsatisfied goals are more richly indexed in

memory upon the realization that they will not be satisfied, and are thus easier to recall. Unsatisfied goals need not be viewed as inherently different from other sorts of mental entities in this case. If, however, the increased memorability of unsatisfied goals is not due to indexing at the time of formation or as a result of frustration, but rather to some inherent property -- to an increased level of arousal or activity compared with satisfied goals, as it were -- then we must conclude that unsatisfied goals are quite different from other sorts of mental entities.

In fact, Zeigarnik went on to perform additional experiments which distinguish these two alternatives, although she did not frame the issue in exactly this way. If unsatisfied goals are more memorable simply because they are more richly indexed upon being frustrated, then it must be the case that such additional indexing occurs as a result of being interrupted. Thus, interrupted tasks -- whether or not still pending -- would presumably be more richly indexed than uninterrupted tasks, simply because of the interruption. However, Zeigarnik found that if a subject was interrupted in the performance of some task, but then subsequently allowed to complete the task after several intervening tasks, that task was no more memorable than any other completed task, i.e., than those which were completed without interruption.

These results seem to indicate that the enhanced memorability of unsatisfied goals does not stem, at least entirely, from richer indexing when they are formed or interrupted. Rather, the processes of retrieval from memory, or of maintenance in memory, must somehow be sensitive to the fact that an unsatisfied goal is unsatisfied. That is, the increased memorability of pending goals reflects some extra effort that is expended on their behalf simply by virtue of the fact that they are pending goals -- it reflects, in other words, an inherent property of pending goals. It might be, for example, that the indices to pending goals are maintained better in memory, while those to satisfied goals are allowed to fade. Although this explanation involves indexing, it is not based on indexing alone: Particular effort must be devoted to maintaining unsatisfied goals in memory. Alternatively, it may be that retrieval is sensitive to the degree of "arousal" of a goal, and that pending goals are more highly aroused. On either account, pending goals are allocated special resources, and they are allocated those resources on a continuing basis.

Now, it might in fact still seem possible to maintain that the greater memorability of pending as opposed to satisfied goals does not entail devoting special effort to pending goals. Rather, it could be argued that the difference is due to special effort devoted upon the satisfaction of a goal -- namely, in the removal of indices that point to the satisfied goal. However, all of the indices to a satisfied goal cannot be removed, since otherwise we would

not be able to remember them at all. Thus, such an approach raises the question of exactly which indices are to be removed. The most obvious answer to this question is to postulate the existence of "pending goal" lists in memory that contain indices to pending goals, so that it is the removal of such indices upon the satisfaction of a goal that explains its decreased memorability as compared to pending goals. What I now want to argue is that this explanation for the Zeigarnik effect is not an alternative to an explanation based on the devotion of special efforts to pending goals, but is rather one way in which such special efforts could be implemented.

First of all, it should be clear that the sort of pending goal list which is capable of accounting for Zeigarnik's results cannot simply be a goal agenda of the sort employed in standard planning models -- that is, a list of the goals and subgoals governing a planner's current behavior -- since many or even most unsatisfied goals are not in fact governing a planner's current behavior. Thus, it must include unsatisfied goals which are not currently being pursued by the agent. The question that then arises, however, is what the purpose of such a pending goal list would be. For, after all, it cannot simply exist so that subjects can more easily report their unsatisfied goals to an experimenter: It must somehow play a role in carrying out the functions that motivate the increased memorability of such goals in the first place. *It must, in other words, play a role in facilitating the recognition of opportunities to* pursue unsatisfied goals. Thus, for example, the inclusion of some goal on a pending goal list in memory cannot simply be for the purpose of indicating that it is pending, for that could be accomplished simply by individually marking the goal. In other words, it makes no sense that such a list merely be used to determine whether or not a goal is unsatisfied after it has been retrieved by other means: Such a scheme neither fulfills the functional requirements of opportunistic planning, nor accounts for Zeigarnik's results.

Thus, the presence of some goal on a pending goal list must actually facilitate the activation or retrieval of the goal when an opportunity arises. That could be accomplished, for example, by using the list itself to index to unsatisfied goals. On such an account of goal activation in opportunistic planning, a certain amount of effort would always be devoted to checking the goals on a pending goal list. Or, it might be that goals on such a list would be maintained in a heightened state of arousal. The point here is that any account of the Zeigarnik effect that relates it to the requirements of opportunistic planning will be functionally indistinguishable from the notion that pending goals are inherently endowed with certain capacities -- that pending goals have memory processing devoted to their retrieval under appropriate circumstances beyond what is devoted to other mental entities, and in particular, to goals that have already been satisfied. Whether the locus of this processing capacity lies in the

goals themselves or in some central processor -- whether, in other words, memory is best viewed as a multi processor or as a time shared system -- is a subsidiary question. The fact that two goals can interfere with each other's performance, however, resulting in bungled actions or conflations, seems to indicate that some kind of parallel processing is involved in intentional behavior in humans (see Norman, 1981).

4. How much processing power?

Having examined both functional arguments and empirical evidence in favor of endowing goals with inherent processing power, we now turn to the thorny issue of how much power they actually need. If, ultimately, we ascribe too much ability, too much intelligence to goals in the service of recognizing and dealing with opportunities for and threats to their achievement, then we will have committed the homuncular fallacy -- "explaining" intelligence in terms of intelligent components. We cannot explain planning, in this case, by postulating that goals can plan -- unless the kind of planning that goals do is simpler than the planning of the agent as a whole.

Thus, any theory which attempts to view the mind in terms of component agents must take the threat posed by the homuncular fallacy very seriously. As a result, there is great pressure to minimize the capabilities of the agents in terms of such a theory attempts to explain the mental abilities of an organism as a whole. In Minsky and Papert's "society" theory of the mind (Minsky, 1979 and 1980), for example, the individual agents appear to be quite simple. They do not communicate with each other via complex representations, for that requires the ability to interpret such representations, which in turn requires that each agent itself be a fairly sophisticated computational device. Instead, Minsky and Papert propose that the individual cognitive agents communicate with each other via direct links, and possess only a limited capacity to construct and interpret symbolic representations.

This is, it seems to me, an inadequate conception of mental agents for the purposes of opportunistic goal processing. We have seen that goals, construed as active mental agents, must be able to infer opportunities for their achievement from features of the situation in which the agent as a whole finds itself, and that, further, whether or not those features actually signify an opportunity is dependent on the context. This context-dependence is a key problem, for it seems to require fairly sophisticated inferential processing to determine whether or not an opportunity exists. For example, consider again Freud's example in which a speaker's goal to insult or ridicule his boss gives rise to the utterance "Gentlemen, I call upon you to hiccough to the health of our chief," in which the word "hiccough" has been substituted for "drink." Recall

that in our explanation of this slip we argued that the word [illegible] during the attempt to retrieve the word 'drink' [illegible] German 'aufzustossen' as opposed to 'anzustossen'. We argued further that such a substitution would only serve to insult or ridicule the boss because of the context – the fact that the utterance was intended as a toast. Furthermore, the effect depends not on the context but on the meaning of the material that has been made available by [illegible] as well. Suppose 'aufzustossen' meant [illegible] Gentlemen, I call upon you to rise to the health of our chief, [illegible] insult or ridiculous, instantaneously.

Thus if this slip really did constitute an attempt to satisfy a goal to insult or ridicule the boss, that attempt involved a very sophisticated evaluation of how an [illegible] substitution would be, an evaluation that depended on both the potential affix substitution and the context in which the substitution would occur. Even without spelling out in detail the sort of knowledge that went into this judgment – knowledge of what constitutes insult or ridicule – it seems clear that it is rather highly developed and quite detailed. The application of such knowledge certainly involves quite a bit of inference, and if that knowledge is at all general – i.e., if our judgment of what constitutes an insult is not merely a list of special cases – then that inference involves *the application of general rules, rules with variables in them*. The use of such rules requires, then, the ability to manipulate variable bindings – which means that under some circumstances at least pending goals are allocated a substantial capacity for general purpose inference.

Moreover, consider the nature of this particular opportunity. It was, inherently, a symbolic structure – the meaning of a sentence in German – that required interpretation by the goal to insult or ridicule the boss. Thus, in order to explain slips of this complexity as the result of opportunistic planning, we must accept that goals, construed as active cognitive agents, can in fact interpret symbolic representations and employ general inference rules. This is not to say that such a conception of goals cannot, in turn, be implemented in terms of simpler agents which do not require such abilities. But the goals themselves must be more complex entities.

5. Unanswered questions as active goals

I have argued above that goals are active cognitive agents, active in the sense that they ask questions about the world, questions about the presence of features that might threaten or facilitate their achievement. Indeed, questions themselves are goals of a sort, and like

unsatisfied goals, unanswered questions seem to remain poised in memory, searching for an answer. One is often reminded of an unanswered question when a situation which seems to provide an answer presents itself. Moreover, just as with the recognition of an opportunity, the realization that a situation presents such an answer may not be immediately obvious -- that is, it may require inference.

Consider the following personal experience. Years ago, while at summer camp, I received a letter from my parents, quite ordinary except that it was postmarked in New York City, while my parents lived in Maryland. This would have made sense if my parents had been visiting New York -- I have many relatives there -- but the letter itself made no mention of this. It seemed odd to me that my parents would send me a letter while visiting New York and not say so in the letter, and this led me to doubt that they were in New York at all.

I then formed all sorts of silly hypotheses to explain the postmark. For example, I thought that perhaps my parents might have mailed the letter to my relatives in New York, and that they in turn mailed it to me at camp -- but what was the point of that? I also thought that maybe the post office had failed to postmark it properly at its origin, and that this had been noticed in New York, where a postmark had then been affixed. For some reason, I don't recall having formed more sensible explanations -- e.g., that my parents had written the letter in Maryland, but hadn't had a chance to mail it before travelling to New York. In the end, then, several questions remained outstanding in my mind. Had my parents been in New York? If so, why hadn't they said so in the letter? If not, how else could it have been postmarked as it was? Because I could not find an adequate explanation for the anomalous postmark, questions persisted.

A few weeks later, I saw one of my cousins from Brooklyn at a neighboring summer camp, and he told me that his father had died the previous month. Immediately, I was reminded of the letter my parents had mailed me. Obviously, they had been in New York for my uncle's funeral, and had mailed the letter to me while there for that reason. Of course, the question then arose of why my parents had not told me about my uncle's death. But what is of interest here is that I was immediately reminded of the heretofore unexplained, and still therefore outstanding, question of how the letter came to be postmarked in New York -- and in particular of the question of whether my parents had been there -- upon finding myself in a situation which seemed to present an answer. How did this take place?

Let's first simply sketch out the inferences that were required to answer the question without making any commitment as to how they were made. Knowing that my uncle had died,

coupled with general knowledge about deaths, led me to infer that he had had a funeral. Knowledge of funerals told me that relatives could be expected to attend, and that it could be expected to occur near the residence of the deceased. Since my parents were relatives of my uncle, and since he lived in New York City, it could be concluded that my parents attended a funeral in or near New York City. Thus, one could finally conclude that they were in New York.

Now, as I have laid out this chain of inferences, I have portrayed it as simple forward chaining from the input. In the case of some of these inferences, such a bottom-up approach seems reasonably plausible. For example, one could imagine that the instantiation of a funeral script was bottom-up, a normal part of understanding the episode of my uncle's death: Hearing that he had died led me to think about his funeral. Perhaps it is even possible that such bottom-up inference extended to filling in such details as the facts that my parents would attend, and that it would be in New York, although I don't recollect thinking about the other relatives who must have attended.

However, it seems to me rather farfetched that all of the crucial inferences in this chain of reasoning were produced by forward chaining in a bottom-up fashion. For example, it seems difficult to motivate the inference that, since my parents were most likely at the funeral, and since the funeral was most likely in New York, my parents must have been in New York, because many other questions and inferences about my uncle's death seem intrinsically more significant. When did my uncle die? What caused his death? How should I comfort my cousin? How was my aunt? All of these questions and their answers were far more significant, really, than the issue of whether or not my parents had been in New York. Yet these questions did not immediately come to mind. Instead, I focussed on my parents' presence in New York, something which is *a priori* much less significant, and which only gained significance because of its relation to the unexplained mystery of the postmark. The problem is, there could be no way of coming to see its significance until after some connection with this unanswered question had been grasped.

It is this precisely this issue -- that the significance of a feature can only be determined by its relation to a goal -- that led us to question the coherence of purely bottom-up approaches to understanding in earlier chapters. In this case, clearly, it seems far more likely that the inference that my parents had been in New York involved the top-down influence of the unanswered, and still pending, question of whether or not they had been there. In other words, some cognitive and inferential capacity must have been allocated to this unanswered question, just as to an unsatisfied goal. How much inferential capacity must we attribute to the

unanswered question in order to construct the appropriate inference chain in this case? First of all, notice that it requires the ability to apply general implication rules, that is, rules with variables in them. We must be able to connect up the question, "Were my parents in New York?" with the assertions "My parents attended my uncle's funeral," and "My uncle's funeral was in New York." This entails the use of a general rule, something like "If someone takes part in some script which occurs in a given location, then they must be at the given location." Thus, the unanswered question must be provided with the inferential resources necessary to examine and interpret representations, and to draw inferences using general rules.

To sum up so far, I have argued that the inference that my parents must have been in New York was actually of relatively little importance in the events surrounding my uncle's death, and that whatever importance it had was due to the questions raised by the letter that I received from them. Thus, any metric of importance based solely on bottom-up considerations would count this inference as rather minor, and any inference scheme which allocated inferential resources in conformance with such a measure of importance would be unlikely to draw this particular inference. The fact that I nevertheless drew this inference suggests, therefore, that the allocation of inferential resources must be based on other considerations, in particular the relationship between this inference and the pending question. That is, the allocation of inferential resources, as we have already argued, must be due to the unsatisfied goals which reside in memory. Now we come to the crux of the matter: How can the allocation of inferential resources to the features of a situation be made sensitive to the goals which must determine that allocation?

One answer -- the most direct -- is simply to allocate sophisticated computational resources to every goal in memory and use back-chaining as the preferred method of inference. In such a "brute force" top-down approach, the detection, representation, and evaluation of significant features of a situation would all be accomplished by means of checking from the goals themselves. This seems incredibly expensive, and hence implausible -- but no more so than equally "brute force" bottom-up approaches. If anything, I suspect that the latter is more expensive -- that more inferences can be drawn from inputs on average than from an agent's goals, or to put this another way, that the branching factor is much worse bottom-up than top-down. Think of all of the potentially significant inferences that might be drawn from a single visual scene -- for example, the relative distances between any two pairs of points.

Still, a more integrated approach to the allocation of inferential resources -- one that takes into account both the agent's pending goals and the situation in which he finds himself -- would be preferable. For example, the two-tier model of opportunity recognition proposed in

the last chapter allocates inferential resources in part, on the basis of features that are both relatively easy to compute and capable of providing an approximate estimate of the relevance of pending goals with respect to a situation — or, equivalently, the significance of features of that situation with respect to pending goals. Such an approach might be applied to the processing of unanswered questions as well.

Let's see how that might be accomplished in this case. One of the unanswered queries concerned my parents' presence in New York. If the bottom-up instantiation of the *funeral* script would indeed have involved assigning my parents and New York as values of its **Mourners** and **Location** roles respectively, then it is possible that a parallel indexing method such as spreading activation or marker passing would have aroused this query, since it too mentions both my parents and New York. Perhaps even a reference to my parents alone would have done the trick. Then, upon the arousal of the query, inferential effort could have been allocated to ascertaining whether the situation actually contained any evidence relevant to its answer. In other words, inference would only be employed *if* there were evidence that it might be useful.

However, it must be emphasized again that even under the two-tier model goals must be construed as active entities. *Many memories* could be retrieved even using both "my parents" and "New York" as indices: my parents grew up in New York, they met in New York, they were married in New York, and they often visited New York. Even more detailed knowledge was available to me concerning trips we took to New York together, things that we did there, etc. Thus, something else is needed to account for my ability to specifically pick out the query "Were my parents in New York recently?"

This brings us to one of the most interesting aspects of this example. The original mystery of the postmark posed several questions, one of which was the following: Why wouldn't my parents tell me if they had been in New York? As I pointed out above, my cousin's report of his father's death posed a similar riddle: Why didn't I already know that my uncle was dead? Why hadn't my parents told me? Thus, both of these situations raised similar questions about why my parents wouldn't have told me something that I would have expected them to tell me. Assuming then that I had considered the question of why I didn't already know about my uncle's death, it seems plausible that I would have been reminded of the earlier mystery.

6. Conclusion

In the last chapter we saw that [illegible] conception of goals as active cognitive agents. However, [illegible] ascribed to goals remains an open question. [illegible] themselves to the attention of the agent as a whole [illegible] constitute opportunities for their pursuit. However, to the extent [illegible] features requires inference, there is a functional basis [illegible] resources to pending goals, at least under some circumstances.

In this chapter, I have presented some empirical evidence [illegible] as dynamic, active mental entities, and explored further the [illegible] view. We saw, first, that the intentional [illegible] of the tongue could be functionally justified on the grounds that [illegible] opportunistic planning. This analysis led us to the conclusion that, in an opportunistic planner, no unsatisfied goal is ever really suppressed. We saw, further, that in some cases the intentional explanation of a slip entails the attribution of inferential resources to the recognition and pursuit of opportunities to achieve suppressed goals. This implies that under some circumstances *the degree of activity of an unsatisfied, pending goal is very high indeed.*

Further empirical evidence for the view that goals are active mental agents can be found in Zeigarnik's experimental results, which show that pending goals are more memorable than satisfied goals. I also presented anecdotal evidence that unanswered questions, like pending goals, seem to remain active in memory, awaiting situations which might provide an answer. Our analysis of one such example seemed to indicate the need for highly sophisticated inferential processing, including in particular the ability to use general rules and manipulate variable bindings.

CHAPTER 10

CONCLUSIONS

1 Functional arguments and artificial intelligence

Throughout this thesis I have referred to [illegible] methodological guidelines which seem useful [illegible] intelligence. The particular problem which I have [illegible] computational model be an explanation of some aspect of [illegible] be judged? In the absence of any such criteria, the only standard that exists is the [illegible] behavior of a system. Yet, it has been clear at least since the appearance of Weizenbaum's (1966) famous ELIZA program that there is something seriously amiss with such a conception. Nevertheless, it remains in many cases the default standard. The notion that a computer implementation constitutes evidence for the sufficiency of a theory is still widely held, and such arguments almost always come down to this line: *The program implements the theory*, and the program has the following (impressive) input output behavior.

But an AI program does not, in most cases, completely implement the theory that is proposed -- indeed, in most cases, the theory itself is seriously incomplete. The metaphor that I think applies best is that of a building under construction without complete plans. Some of the walls are up, some of the foundation has been laid, but a lot of scaffolding surrounds the project and many temporary load-bearing beams are helping to hold the incomplete structure together. The scaffolding and temporary beams are the inescapable hacks that are necessary to get the program to work on even one or two examples. There is no way of knowing, really, what is responsible for the behavior of the program. Is it due, mostly, to the structure that is under construction, to the designer's incomplete conception -- i.e., to those elements of the program that implement, in part, the partial theory -- or is it due to the scaffolding and temporary beams that surround and support the program -- i.e., to the hacks?

There is only one way in which this uncertainty could be overcome, and that would be to argue, for each behavior or ability of interest, that the program's performance followed from the theory, not from the temporary scaffolding. Such arguments are difficult to find in the literature, and they are difficult to construct. The behavior of a complex program and the

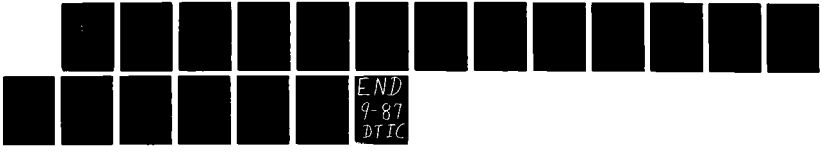
relationship of the program's structure to that behavior are notoriously difficult to understand. But there is no choice. In the absence of such arguments, there is simply no reason to believe that a program's performance in any way reflects the theory that it purports to implement. Indeed, if the program itself constitutes the bulk of the work that is being reported, then in the absence of such an argument, there may well be no explanation of the behavior or ability at all.

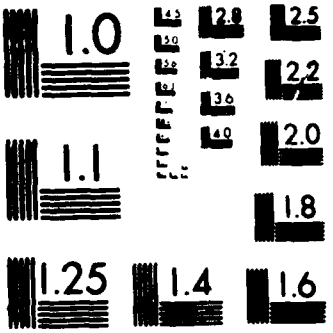
We have seen many examples of such problems in earlier chapters. Let me give just one more hypothetical example here, one that I hope suggests how we might overcome this methodological quagmire. Consider the following "process model" of Freudian slips: We start by gathering a large corpus of slips. Then, we analyze it to yield statistical patterns -- how often one word is substituted for another. We might, further, analyze the data for conditional probabilities -- how often one word is substituted for another in the context of a third -- and thus gain some statistical knowledge about the role of context. Alternatively, we might start by conjecturing that word associations are causally implicated in slips, so that the probability that one word will be substituted for another is dependent on the strength of the association between them. In that case, rather than gathering a corpus of slips directly, we would gather information about word associations using a variety of experimental techniques. Again, we might also gather data about the context-sensitivity of word associations, i.e., the strength of *the association between two words in the context of a third.* On the basis of our original hypothesis, this data could be used to predict the likelihood that one word will be substituted for another in the context of a third.

Now, using this sort of statistical data -- based either on word associations or directly on observed slips -- we construct a device which I will call a "slipper." The slipper consists of a large table representing the probability that one word will be substituted for another -- a "confusion matrix" -- based either on word associations or direct observation. If we have gathered data about the conditional probability of a substitution depending on context, we represent that information in the table as well. For any given word in the table, the highest probability entry is of course the word itself.

We are now ready to construct our process model of Freudian slips. We [illegible] language generator, and direct its output -- a perfectly well-formed [illegible] -- to the "slipper." For one or more words in the utterance, the slipper will randomly substitute another word according to the probabilities in the confusion matrix. Because the highest probability substitution is the identity, the output of the slipper will almost always be identical to the input, but occasionally another word will be substituted and the output will [illegible]

AD-A103 553 INTEGRATED PROCESSING IN PLANNING AND UNDERSTANDING(U) 3/3
YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE
L BIRNBAUM DEC 86 YALEU/CSD/RR-489 N00014-75-C-1111
UNCLASSIFIED F/G 12/9 NL





have a "process model" which, in fact, produces behavior which appears to be exactly like Freudian slips. Moreover, if our hypothesis that associations are implicated in the formation of slips is correct, the model will even be predictive: It will produce slips which have not been previously observed, but which will in fact be found to occur, and it will produce them in the correct proportions. However, despite its brilliant success in producing the desired behavior, we cannot help be ask, does this model really *explain* Freudian slips?

The answer is clearly no. The model does not explain slips any more than the statistical information upon which it is based. It simply encodes statistical facts about word associations and slips in a rather complex form, as a program. The fact that this statistical description is implemented in a computer program adds nothing to the description itself, because the "Freudian slipper" is a ludicrous device: It serves no purpose. There is no reason why a model of generation should include a device whose only purpose is to introduce errors into its output. The point is, our intuition that this model fails to provide an adequate explanation of Freudian slips is due to the fact that there is no functional justification for the rules and processes it invokes in order to produce them.

What can be done to avoid this kind of situation? How can we construct a process model which is capable of providing a genuine explanation of such a behavior as Freudian slips? The answers to these questions depend on the notion of *functionality*. Behavior can only be explained by a process theory if, first, it can be shown to arise as a result of the rules and processes which constitute that theory, and second, if those rules and processes can be justified for some reason *other* than simply to give rise to the behavior in question. The necessity for this second requirement stems from the protean nature of computer programs: Obviously, if a given output is specified for some input, or a specific output is to be avoided, a program can always be written so that it produces or blocks the specified output. This is why the notion of functionality is critical. Rules and processes in a process model of some cognitive ability must be shown to be useful in performing that ability. If functional arguments can be advanced as to why the rules and processes are as they are, then those arguments, and the demonstration that those rules and processes give rise to some piece of behavior, constitute an explanation of the behavior. But by themselves, without some functional justification, the rules and processes that produce some behavior are simply an alternative description of that behavior. In sum, if certain rules and processes are put into a theory or program simply to cause it to exhibit some behavior, then, unless that behavior can itself be functionally justified, no explanation of the behavior has been developed.

In this thesis, I have taken the position that a good theory in artificial intelligence is a

functional theory. Thus, the process model of Freudian slips presented in chapter nine, although by no means complete, contrasts sharply with the "slipper" theory concocted above: The explanation of Freudian slips given there derives from processes that are functionally justified by the need to perform real-time opportunistic planning. It seems hard to imagine an alternative form of explanation in the cognitive sciences. Once functional considerations are seen to be paramount in computational theories, it seems to me that the criteria by which such theories are to be judged become much clearer.

2. Functional arguments and integrated processing

The methodological theme of this thesis might be phrased as follows: Programmers should not make arbitrary decisions in the design of process models. The scientific theme is quite similar: Neither should programs. A program that makes arbitrary decisions must, inevitably, make most of them incorrectly, and must, as a result, rely on backtracking. One way to avoid backtracking is simply to avoid presenting the program with any choices to make. In a sense, the program still makes arbitrary choices, but those choices somehow always turn out to be correct. This is the wishful control structure fallacy: What would have been an arbitrary choice of the program is now an arbitrary choice of the programmer. There is, in other words, a clear connection between the need to functionally justify the rules, representations, and processes that make up a model, and the need for the model itself to make justified decisions. Without the methodological constraint imposed by functional requirements, arbitrary decisions by a program can always, in any given instance, be avoided if the programmer simply restricts the options available to the program. But if this restriction is based on nothing more than the need to get the program or model to exhibit the proper behavior in the case at hand, then it is just as arbitrary. It is, indeed, clearly counter to the functional requirements of the task in general.

Thus, in order to ensure that process models actually explain something, we must avoid the wishful control structure fallacy. On the other hand, even if the model itself makes the decisions, if it simply makes them arbitrarily -- if it relies inherently on non-determinsm -- then again, it seems to lack any explanatory power. From the point of view of computational explanation, models that rely crucially and centrally on arbitrary choice and backtracking are just too general: They can be applied anywhere that a good descriptive theory has been developed. On the more practical side, the inefficiency of non-determinism is well-known. Indeed, if non-determinism were not such a problem, it could be argued that theorem-proving methods would have proven far more widely applicable to AI problems than has in fact been the case.

So, if computational models are to have any explanatory force, we must search for some middle ground: Programs must make their own decisions, but they must not make them arbitrarily if that can be avoided. As I argued in chapter one, the way to make rational decisions -- as opposed to arbitrary ones -- is to make informed decisions. Integrated processing is motivated by this need to bring as much relevant information to bear as possible to bear on decision-making. That is, once it is accepted that the need to make motivated decisions is a central explanatory requirement on any process model, the researcher's task must include an attempt to specify what sort of information is relevant to what sort of decision, and to examine whether and how that information can be made available to the decision process, and to show how it can be taken into account.

Once this task is undertaken, there are two possible outcomes: In the simplest cases, it may be possible to ascertain exactly what evidence can be expected to be present, and how it should weigh in the decision. Realistically, this case does not seem to arise very often. For example, it might be thought that the information necessary to correctly disambiguate a word would be clearly specifiable ahead of time, so that a decision rule could be indexed with the word to check for the presence or absence of certain features. As we have seen however, this is not the case. Instead, we are confronted with an instance of the second possible outcome, namely, that exactly what information will be relevant to making a decision cannot be specified in advance. That is, the decision process for lexical disambiguation must be capable of taking into account diverse evidence, and for any given decision, exactly what that evidence will be cannot be specified in advance. This is not to say that the situation is entirely hopeless: Although we cannot hope to specify exactly what information will be required, we can still arrive at an understanding of what sort of information seems to influence the decision, and how that information can be brought to bear. The fact remains, however, that the decision process for this task must be capable of combining facts not specified in advance to yield a rational conclusion as to the best way -- given those facts -- to make the decision.

This ability to combine evidence that has not been completely specified in advance seems to be a hallmark of inferential reasoning: It seems to defy representation in a simple table or static decision tree. For example, it is apparent that all attempts to perform lexical disambiguation using such non-inferential techniques are inherently inadequate. But inferential processes seem expensive, and they are poorly understood. Thus, the objection may be raised that the cost of making rational decisions by means of inference may turn out to be higher than simply plunging ahead and backtracking if mistaken.

Although this objection to integrated processing may turn out to be a valid, we can only

hope that it is not -- for, if true, there will be little of interest to say about mental processing beyond a description of the domain knowledge that is necessary in order to behave intelligently. Moreover, I believe that this position reflects an unwarranted pessimism about our ability to come to grips with the problems of inference. It cannot be denied that sophisticated inferential processing must ultimately play a central role in any model of intelligent behavior. Given that we know it will be necessary at some point, constantly hobbling our theories of planning, understanding, and learning by attempting to avoid its use at all costs reflects a misplaced sense of parsimony. It is rather as if one were designing some system that one knew ahead of time required a computer to fulfill its primary functions, but then, in all sorts of places where the computer might be useful or even essential, one struggled to design solutions that did not make use of it. The point is that parsimony arguments cannot be applied to arbitrarily isolated components of a goal-directed system: They must be applied to the system as a whole.

I realize that the point of view I am criticizing here stems from a valid fear of falling into the homuncular fallacy, whereby some unanalyzed subprocess -- in this case "inference" -- is doing all the work, and the rest of the model is just window dressing. The way around this is not to shun inference, however: That only raises the prospect that the entire model will turn out to be window dressing. Rather, we must try and specify the inference processes that will be required for the given task, and the sort of information that it must take into account. We must, in other words, attempt to discover the functional constraints that apply to the problem. That is what I have tried to do in this thesis.

3. Top-down and bottom-up

One issue that I think this work has shown to be misconceived is the controversy over whether understanding is "top-down" or "bottom-up." The real issue, it should now be clear, is whether or not it is integrated. An integrated model of understanding is not purely top-down, but it is relatively top-down compared with a non-integrated model. This follows from the fact that in order to achieve the goal of avoiding backtracking, as much relevant information as possible must be brought to bear on decisions that arise early in the process of understanding. Such information includes, in particular, the goals and hypotheses of the understander. Thus, the underlying functional motivations of integrated processing impose strong pressure in favor of understanding models which are relatively top-down.

When the focus shifts to the issue of planning, the result of integration is entirely different. An integrated model of planning will be, relative to a non-integrated model, rather

"bottom-up." That is, decisions about which goals and plans to pursue will be based, in part, on the situation in which the planner finds itself, rather than being governed solely by higher level goals and plans. Moreover, as in the case of understanding, the sooner that these relevant contextual factors can be taken into account in planning, the better from the point of view of avoiding backtracking. Thus, the underlying functional motivations of integrated processing impose strong pressure in favor of planning models which are relatively bottom-up.

In a sense, this disparity in the implications that integrated processing has for planning, on the one hand, and for understanding, on the other, should be obvious. After all, if one takes a strictly top-down model of planning, and turns it around, one has a strictly bottom-up model of understanding. Such models, clearly, do not share a commitment to either top-down or bottom-up processing. Rather, what they have in common is a commitment to a strict division of processing into separate modules, each defined solely in terms of its access to a single source of information, and arranged, logically and temporally, in a simple linear sequence. So the issue is not whether understanding should be top-down, or planning should be bottom-up. The issue is whether both planning and understanding should be integrated in order to avoid the need to make arbitrary decisions.

Now, I make this point in part because I suspect that many cognitive scientists will be ready to accept as unproblematic the notion that planning should be integrated, and hence partly bottom-up -- including many who will be bothered by the idea that understanding should be partly top-down. My challenge to them is: What is the difference? The functional arguments underlying a relatively bottom-up approach to planning -- the need for integrated processing to avoid arbitrary decisions -- seem to apply just as well to the case of understanding. And when applied to understanding, integrated processing leads to relatively top-down models. It seems to me, therefore, that those who believe that understanding is a strictly bottom-up process must take one of two positions: Either they must argue that planning and understanding are so fundamentally different that, although the arguments in favor of integrated processing apply to one, they do not apply to the other, or else they must reject the view that planning should be an integrated process, and, in so doing, embrace the position that it is entirely top-down. Both propositions strike me as dubious.

4. Integrated understanding

As I said above, an integrated approach to understanding leads to models which are relatively top-down. In part I of this thesis, I sketched out some of the consequences of this view, gave some of the evidence in favor of it, and criticized non-integrated theories. In this

section I would like to review, briefly, what I said there.

In chapter two, I started by investigating the implications of integrated processing for the relationship between syntax, semantics, and pragmatics in language understanding. I argued that, on an integrated view, although purely syntactic rules exist and play a role in language understanding, there is no need to suppose that they are applied by an independent process, or even that operate on independent, purely syntactic representations. Thus, on this view syntax plays no privileged role in language understanding. Syntactic evidence and syntactic knowledge are used, to be sure, but so is other knowledge, and more to the point, other sorts of knowledge do not need to "wait their turn" while syntactic processing takes primacy. Nor, as in such theories as those proposed by Fodor, Bever, and Garrett (1974), or Marcus (1980), or their intellectual heirs, is semantics to be viewed as a "subroutine" employed by syntax to make the tough calls. This notion seems to represent, if anything, an inversion of priorities.

In this chapter, I also addressed one of the key problems facing an integrated approach to understanding, namely, the need to flexibly combine knowledge from different sources in order to make a decision. In particular, I discussed how lexical and syntactic information, on the one hand, and specific constraints arising from detailed memory structures, on the other, could be combined dynamically to create expectations that could brought to bear on the lexical level, and which could therefore be used for such tasks as lexical disambiguation. In the context of this problem, I also developed the distinction between vagueness and ambiguity, and showed that an integrated approach to understanding -- because it is not bound by the view that the meaning of a sentence is simply a combination of the meanings of the words that make it up -- leads to a simpler and more general treatment of the interpretation of vague words than do previous approaches.

In chapter three, I developed a critique of the largest class of non-integrated theories of understanding, those concerned with modular syntactic analysis. My critique was based on the failure of these theories to address the kinds of functional considerations that are the primary concern of artificial intelligence. I pointed out, first, that these theories fail to justify the form of the output which they produce, and in fact often make no claims whatsoever about their output. I also criticized the inherent reliance on non-determinism by ATNs and Prolog-based parsers, and argued that such control mechanisms are totally lacking in empirical content. The only claims made by such parsers, really, are embodied in the grammars which they employ -- typically taken directly from the linguistics literature. I also showed that Marcus's more recent deterministic theory of syntactic parsing, by its failure to address such problems as lexical ambiguity and genuine structural ambiguity, also fails to assert or support an empirically

meaningful claim of syntactic modularity. By ignoring all of the problems in language analysis which seem to require the heavy use of semantics, this theory, it seems to me, is simply begging the question.

In chapter four, I turned to an analysis of the problem of lexical ambiguity from an integrated point of view. Reviewing first the disambiguation methods employed by syntactic analyzers, we saw that they either depend on backtracking, or else simply fail to address the issue. I then turned to a review of semantic and conceptual methods for lexical disambiguation, primarily selectional restrictions and scriptal lexicons. We saw that such methods cannot work in general, and cannot work because they attempt to avoid the need to bring inference to bear on the problem: Lexical disambiguation, as has been argued since Bar-Hillel (1960), requires plausible inference taking into account both context and arbitrarily complex world knowledge -- in particular, it requires the ability to perform abduction, or inference to the best explanation for an input. Since lexical ambiguity must be resolved quickly in order to avoid the need for non-determinism, the fact that its resolution requires context-based inference forms one of the chief functional arguments in favor of an integrated approach to language analysis. Nevertheless, we saw that many putatively integrated models of language understanding in fact fail to go beyond the traditional methods of selectional restrictions and scriptal lexicons, and do not employ inference in the resolution of lexical ambiguity or other problems in language analysis. In other words, these models simply fail to address the issues which motivate integrated processing in the first place. More positively, I argued that the problem of lexical ambiguity imposes important requirements on the process of explanatory inference. In particular, we saw that explanatory inference rules cannot be expected to attend to all of the features of a situation that might affect the applicability of the explanations which they offer, and that more general criteria for judging the acceptability of an explanation are therefore necessary.

In chapter five, we discussed the implications of an integrated approach for the inferential memory processing involved in explanation-based understanding, particularly in light of the need to employ abstract thematic knowledge as proposed by Schank (1982) and Lehnert (1981). Integrated understanding entails the use of possible hypotheses about a situation to limit the inferences which are drawn from inputs in that situation. The utility of such an approach is easily demonstrated by the advantages of a very simple model, bi-directional inference from both inputs and hypotheses. However, this model still entails a great deal of wasted inference, especially if a large number of inferences can be drawn from any given proposition, as seems to be the case in highly familiar domains. In view of this problem, I specified the relevant functional criteria for a more highly integrated theory of

explanatory inference in understanding. We saw that there is one model which meets these criteria, script/frame theory. However, the limitations of script/frame theory are well known. The main problem is the utter inflexibility of script-based inference: Expectations are carefully tuned to exactly and only a limited class of inputs. In other words, exactly the property which makes these models highly integrated is also what makes them too inflexible. We thus sought a model which would still be integrated, but less inflexible.

Whatever the solution, it seems clear that if an integrated approach is to be feasible -- that is, if an hypothesized explanation is to play a useful role in directing the inference process which attempts to explain an input -- then it must be because the knowledge associated with the hypothesis can be employed to provide such direction. The problem is that this knowledge may not be expressed in terms commensurate with the input, particularly if it is highly abstract, as it is in the case of thematic knowledge. In script/frame theory, of course, this is guaranteed: Expectations -- or, if one prefers, "active memory structures" -- are couched in exactly the representational vocabulary that input is expected to be couched in. But this is exactly what gives rise to the inflexibility which must be overcome in a more general theory. Thus, the integrated use of hypotheses in directing inference will, in general, involve a "translation" process whereby the input is represented in terms commensurate with the hypothesis. Our problem, then, is to perform this translation process in a way which does not, itself, depend on undirected inference. I proposed a solution to this problem based on the use of simple indexing techniques to guide the inference necessary to transform an input into terms commensurate with an hypothesis.

It should be clear that my specific proposals are only a first step towards an integrated theory of understanding: They are both overly simplistic and lacking in many details. What is important, however, is that the motivations behind an integrated theory of understanding, and the criteria which must be met by such a theory, have been worked out to the point where it is possible to see how previous proposals fall short, and to make new proposals that, whatever their shortcomings, at least fall into the class of genuinely integrated theories.

5. Integrated planning

In part II of this thesis, I turned to the question of how integrated processing affects our view of planning. I argued that the need to take external factors into account in planning as early as possible -- which is motivated by the goal of making rational decisions and avoiding blind search -- leads, ultimately, to a model of planning in which goals are often set on the basis of opportunities provided by the situation in which the planner finds itself. In other

words, an integrated model of planning will be an *opportunistic* planner.

In chapter seven, I pursued this argument within the context of conversational planning, particularly in arguments. I showed that previous approaches to conversational planning are too top-down, and must therefore rely on backup to produce workable plans. I then presented several examples in which the memory and inference necessary in order to understand one's interlocutor could be expected to uncover the seeds of a reasonable response. This led, then, to a notion of planning in which such contextual features could play a role in the selection of a conversational goal or plan to be pursued. More specifically, I argued that opportunistic planning seems to account for a fundamental feature of conversational behavior, namely, that conversations seem at one and the same time to be goal-directed and yet rather wandering and disorganized.

After this initial foray into opportunistic planning in the domain of conversational behavior, in chapter eight we turned to a discussion of the problems posed by the approach in general. The chief problem of opportunistic planning is simply noticing that an opportunity exists. An opportunity exists when some feature of the situation in which the planner finds itself facilitates the achievement of a goal not necessarily governing his immediate behavior in that situation. Thus, recognizing an opportunity will, in general, entail recognizing the presence of features which are not necessarily related to the goals governing an agent's current behavior.

Opportunistic goal activation entails indexing goals in terms of features which indicate opportunities for their pursuit, generally unmet preconditions of plans for those goals. In the simplest model -- the "mental notes" model -- there is a central planner of some sort, and goals must be brought to the attention of this central planner when opportunities for their achievement are detected. In the very simplest case, a goal plays no role in the perceptual and inferential processes which are involved in detecting the feature which constitutes an opportunity for its pursuit. Those processes must, therefore, be pursued for some other reason: The detection of features which indicate an opportunity for a goal is entirely bottom-up as far as the goal itself is concerned. However, whether or not the detection of an opportunity involves the top-down influence of the goal itself, any model of opportunistic planning entails characterizing goals as active mental entities. At the very least, they must be able to call attention to themselves -- that is, to place themselves before the consideration of the central planner -- when an opportunity is detected.

Moreover, the features which might usefully indicate the presence of an opportunity to

pursue some goal, but which could conceivably be noticed without the intervention of the goal itself, are very likely to produce many "false alarms." Establishing that an opportunity actually exists will entail further inference in such cases. For example, if the planner's goal is to purchase a particular book, and he forms the plan to buy it at the Yale Co-op, then "Yale Co-op" is a feature which plausibly indicates an opportunity to achieve the goal. However, even if this feature is present, an opportunity to achieve the goal only exists if the planner or some agent of the planner's is actually in the vicinity of the Yale Co-op, if it is open, and if the planner or his agent has sufficient funds. Seeing a story about the Yale Co-op on television, for example, would not constitute an opportunity. We are thus led to a two-stage model of opportunity recognition, in which goals are conceived as active mental entities, capable of noticing the presence of features which indicate potential opportunities -- although not necessarily of playing any role in the detection of those features -- and of calling on further inferential processing when such features are detected in order to determine whether or not a *bona fide* opportunity exists.

There is an additional problem however. Active processing often seems necessary merely in order to detect the features which might indicate an opportunity. This raises two separate but related issues: First, can simple, non-inferential methods be devised to detect the likely presence of features which indicate opportunitiest? And second, if not -- that is, if inference is required -- then is the goal for which those features constitute an opportunity itself involved, in some top-down fashion, in their detection?

The answer to the first question is undoubtedly yes. I discussed several examples of features -- for example, the kind of ambiguity that is the precondition for a *double entendre*, or being reminded of something that shares no superficial features with the current situation -- which cannot be recognized without active, inferential processing. These are all, roughly speaking, features the components of which cannot be enumerated ahead of time. To the extent that an opportunity depends on this sort of feature, one must either argue that its detection is motivated by some goalunder current pursuit, or that the goal for which it indicates an opportunity is itself actively involved in its detection. To the extent that the first case holds, the mental notes model will suffice. To the extent that it doesn't -- to the extent, in other words, that a system is capable of detecting truly novel opportunities to achieve a goal -- it may not.

In chapter nine, I sketched out the conception of planning and plan execution which is necessary in order to accept Freud's intentional explanations for errors, and argued that this intentional architecture can be functionally justified on the grounds that it fulfills the requirements of real-time opportunistic planning. In particular, we saw that in an opportunistic

planner, no goal is ever really suppressed. Thus, I argued that an opportunistic planner will quite naturally exhibit such behavior as Freudian slips. However, the opportunities seized upon in Freudian slips cannot, in general, be precisely specified in advance. Slips, therefore, reflect an ability to detect and seize *novel* opportunities -- opportunities which are not foreseen or even foreseeable ahead of time. Such an ability seems to entail the allocation of inferential power to the detection of opportunities for suppressed goals. Thus, we were led to the possibility that under some circumstances goals are far more active mental entities than implied even by the mental notes model of opportunistic planning. In this context, we reviewed Zeigarnik's experimental results indicating that pending goals are more memorable than satisfied goals -- results which constitute a converging line of evidence in favor of the proposition that pending goals are active mental agents.

6. Conclusion

In this thesis, I have tried to provide the functional motivations for integrated processing in planning and understanding, to show why such an approach is necessary in order to solve many real problems in artificial intelligence, to characterize some of the problems that arise in attempting to construct integrated models, and to propose some solutions to these problems. The need for an integrated approach to understanding has been argued many times before: Here, I have tried to put together a more coherent account of this view, by articulating and justifying the functional constraints that an integrated approach imposes on models of understanding. Along the way, we have seen how many previous attempts have gone awry, and have as a result gained a clearer view of the problems which must be addressed by a *bona fide* integrated theory of understanding.

In planning, as in understanding, integrated processing is motivated initially by the need to avoid arbitrary decisions and the backtracking that must inevitably follow in the wake of such decisions. But it has larger consequences for our conception of intentional behavior as well. An integrated approach to planning leads to models in which a planner's goals are set in part by the opportunities offered by the situation in which he finds himself -- in other words, to *opportunistic* models of planning. This view raises a whole host of new issues in planning and understanding, probably the most difficult being the problem of how to recognize opportunities when they arise. I have sketched out several possible solutions to this problem, and although much work needs to be done, one thing is clear: This problem must change our conception of what kind of mental construct a goal is. Although the degree of activity may vary, on any account of opportunistic planning goals must be construed as active mental agents. We are led, in other words, to a dynamic conception of goals quite similar to that found in psychoanalysis.

REFERENCES

Abelson, R. 1973. The structure of belief systems. In R. Schank and K. Colby, eds., *Computer Models of Thought and Language*, W. H. Freeman, San Francisco, pp. 287-339.

Ackley, D., Hinton, G., and Sejnowski, T. 1985. A learning algorithm for Boltzmann machines. Cognitive Science, vol. 9, pp. 147169.

Allen, J., and Perrault, C. 1980. Analyzing intention in utterances. *Artificial Intelligence*, vol. 15, pp. 143-178.

Bar-Hillel, Y. 1960. The present status of automatic translation of languages. In F. Alt, ed., *Advances in Computers 1*, Academic Press, New York, pp. 91-163.

Birnbaum, L. 1982. Argument molecules: A functional representation of argument structure. *Proceedings of the 1982 AAAI Conference*, Pittsburgh, PA, pp. 63-65.

Birnbaum, L. In preparation. A functional approach to the representation of arguments.

Birnbaum, L., and Collins, G. 1984. Opportunistic planning and Freudian slips. Proceedings of the Sixth Cognitive Science Conference, Boulder, CO, pp. 124-127.

Birnbaum, L., Flowers, M., and McGuire, R. 1980. Towards an AI model of argumentation. *Proceedings of the 1980 AAAI Conference*, Stanford, CA, pp. 313-315.

Birnbaum, L., and Selfridge, M. 1981. Conceptual analysis of natural language. In R. Schank and C. Riesbeck, eds., *Inside Computer Understanding: Five Programs Plus Miniatures*, Lawrence Erlbaum, Hillsdale, NJ, pp. 318-353.

Bobrow, D., and Fraser, B. 1969. An augmented state transition network analysis procedure. *Proceedings of the First IJCAI*, Washington, DC, pp. 557-567.

Bolinger, D. 1979. Pronouns in discourse. In T. Givon, ed., *Syntax and Semantics, Vol. 12: Discourse and Syntax*, Academic Press, New York, pp. 289-309.

Carbonell, J. 1981. *Subjective Understanding: Computer Models of Belief Systems*. UMI Research Press, Ann Arbor, MI.

Carbonell, J. 1982. Metaphor: An inescapable phenomenon in natural-language comprehension. In W. Lehnert and M. Ringle, eds., *Strategies for Natural Language Processing*, Lawrence Erlbaum, Hillsdale, NJ, pp. 415-434.

Charniak, E. 1972. Towards a model of children's story understanding. Technical report no. 266, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA.

Charniak, E. 1978. On the use of framed knowledge in language comprehension. *Artificial Intelligence*, vol. 11, pp. 225-265.

Charniak, E. 1981. Six topics in search of a parser: An overview of AI language research. *Proceedings of the Seventh IJCAI*, Vancouver, B.C., pp. 1079-1087.

Charniak, E. 1983. Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science*, vol. 7, pp. 171-190.

Charniak, E. In press. Motivation analysis, abductive unification, and non-monotonic equality. To appear in *Artificial Intelligence*.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Clark, H., and Clark, E. 1977. *Psychology and Language: An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich, New York.

Clark, H., and Lucy, P. 1975. Understanding what is meant from what is said: A study in conversationally conveyed requests. *Journal of Verbal Learning and Verbal Behavior*, vol. 14, pp. 56-72.

Cohen, P., and Perrault, C. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, vol. 3, pp. 177-212.

Collins, G. In preparation. Untitled Ph.D. thesis, Yale University, Dept. of Computer Science, New Haven, CT.

Colmerauer, A. 1978. Metamorphosis grammars. In L. Bolc, ed., *Natural Language Communication with Computers*, Springer, Berlin.

Cottrell, G. 1984. A model of lexical access of ambiguous words. *Proceedings of the 1984 AAAI Conference*, Austin, TX, pp. 61-67.

Crain, S., and Steedman, M. 1985. On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, and A. Zwicky, eds., *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Cambridge University Press, Cambridge, England.

Cullingford, R. 1978. Script application: Computer understanding of newspaper stories. Research report no. 116, Yale University, Dept. of Computer Science, New Haven, CT.

Deese, J. 1978. Thought into speech. *American Scientist*, vol. 66, pp. 314-321.

Dehn, N. In preparation. Memory and creativity. Ph.D. thesis, Yale University, Dept. of Computer Science, New Haven, CT.

DeJong, G. 1979. Skimming stories in real time: An experiment in integrated understanding. Research report no. 158, Yale University, Dept. of Computer Science, New Haven, CT.

Dresher, B., and Hornstein, N. 1976. On some supposed contributions of artificial intelligence to the scientific study of language. *Cognition*, vol. 4, pp. 321-398.

Dyer, M. 1983. *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. MIT Press, Cambridge, MA.

Fahlman, S. 1979. *NETL: A System for Representing and Using Real-World Knowledge*. MIT Press, Cambridge, MA.

Fikes, R., Hart, P., and Nilsson, N. 1972. Learning and executing generalized robot plans. *Artificial Intelligence*, vol. 3, pp. 251-288.

Fikes, R., and Nilsson, N. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, vol. 2, pp. 189-208.

Fodor, J., Bever, T., and Garrett, M. 1974. *The Psychology of Language: An Introduction*

to Psycholinguistics and Generative Grammar. McGraw-Hill, New York.

Fong, S., and Berwick, R. 1985. New approaches to parsing conjunctions using Prolog. *Proceedings of the Ninth IJCAI*, Los Angeles, CA, pp. 870-876.

Freud, S. 1935. *A General Introduction to Psychoanalysis*. J. Riviere, trans., Liveright, New York.

Gershman, A. 1979. Knowledge-based parsing. Research report no. 156, Yale University, Dept. of Computer Science, New Haven, CT.

Gibbs, R. 1979. Contextual effects in understanding indirect requests. *Discourse Processes*, vol. 2, pp. 1-10.

Granger, R. 1980. When expectation fails: Towards a self-correcting inference system. *Proceedings of the 1980 AAAI Conference*, Stanford, CA, pp. 301-305.

Granger, R., Eiselt, K., and Holbrook, J. 1986. Parsing with parallelism: A spreading-activation model of inference processing during text understanding. In J. Kolodner and C. Riesbeck, eds., *Experience, Memory, and Reasoning*, Lawrence Erlbaum, Hillsdale, NJ, pp. 227-246.

Grice, H. 1975. Logic and conversation. In P. Cole and J. Morgan, eds., *Syntax and Semantics, Vol. 3: Speech Acts*. Academic Press, New York, pp. 41-58.

Grosz, B. 1979. Utterance and objective: Issues in natural language communication. *Proceedings of the Sixth IJCAI*, Tokyo, pp. 1067-1076.

Hammond, K. 1986. Case-based planning: An integrated theory of planning, learning, and memory. Research report no. 488, Yale University, Dept. of Computer Science, New Haven, CT.

Hayes, P. 1979. The naive physics manifesto. In D. Michie, ed., *Expert Systems in the Micro-Electronic Age*, Edinburgh University Press, Edinburgh.

Hayes, P. 1985. Naive physics I: Ontology for liquids. In J. Hobbs and R. Moore, eds., *Formal Theories of the Commonsense World*, Ablex, Norwood, NJ, pp. 71-107.

Hayes-Roth, B., and Hayes-Roth, F. 1979. A cognitive model of planning. *Cognitive Science*, vol. 1, pp. 395-420.

Hilgard, E. 1956. *Theories of Learning*, 2nd ed. Appleton-Century-Crofts, New York.

Hirst, G., and Charniak, E. 1982. Word sense and case slot disambiguation. *Proceedings of the 1982 AAAI Conference*, Pittsburgh, PA, pp. 95-98.

Hobbs, J. 1979. Conversation as planned behavior. *Proceedings of the Sixth IJCAI*, Tokyo, pp. 390-396.

Hopfield, J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences U.S.A.*, vol. 79, pp. 2554-2558.

Judson, H. 1979. *The Eighth Day of Creation*. Simon and Schuster, New York.

Kaplan, R. 1975. On process models for sentence analysis. In D. Norman and D.

Rumelhart, eds., *Explorations in Cognition*, W. H. Freeman, San Francisco, pp. 117-135.

Katz, J., and Fodor, J. 1963. The structure of a semantic theory. *Language*, vol. 39, pp. 170-210.

Katz, J., and Postal, P. 1964. *An Integrated Theory of Linguistic Descriptions*. MIT Press, Cambridge, MA.

Kolodner, J. 1984. *Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model*. Lawrence Erlbaum, Hillsdale, NJ.

Kuno, S. 1965. The predictive analyzer and a path elimination technique. *Communications of the ACM*, vol. 8, pp. 453-462.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Lalljee, M., and Abelson, R. 1983. The organization of explanations. In M. Hewstone, ed., *Attribution Theory: Social and Functional Extensions*, Blackwell, Oxford, England.

Lebowitz, M. 1980. Generalization and memory in an integrated understanding system. Research report no. 186, Yale University, Dept. of Computer Science, New Haven, CT.

Lehnert, W. 1979. Text processing effects and recall memory. Research report no. 157, Yale University, Dept. of Computer Science, New Haven, CT.

Lehnert, W. 1981. Plot units and narrative summarization. *Cognitive Science*, vol. 5, pp. 293-331.

Levin, J., and Moore, J. 1977. Dialogue-games: Metacommunication structures for natural language interaction. *Cognitive Science*, vol. 1, pp. 395-420.

Levy, D. 1979. Communicative goals and strategies: Between discourse and syntax. In T. Givon, ed., *Syntax and Semantics, Vol. 12: Discourse and Syntax*, Academic Press, New York, pp. 183-210.

Lytinen, S. 1984. The organization of knowledge in a multi-lingual, integrated parser. Research report no. 340, Yale University, Dept. of Computer Science, New Haven, CT.

Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA.

Marcus, M. 1984. Some inadequate theories of human language processing. In T. Bever, J. Carroll, and L. Miller, eds., *Talking Minds: The Study of Language in Cognitive Science*, MIT Press, Cambridge, MA, pp. 253-278.

Marcus, M., Hindle, D., and Fleck, M. 1983. D-theory: Talking about talking about trees. *Proceedings of the 21st ACL Conference*, Cambridge, MA, pp. 129-136.

Marr, D. 1977. Artificial intelligence: A personal view. *Artificial Intelligence*, vol. 9, pp. 37-48.

Marshall, J. 1980. The new organology. *The Behavioral and Brain Sciences*, vol. 3, pp. 23-25.

Marslen-Wilson, W., Tyler, L., and Seidenberg, M. 1978. Sentence processing and the

clause boundary. In W. Levelt and G. Flores d'Arcais, eds., *Studies in the Perception of Language*, John Wiley, Chichester, England, pp. 219-246.

McDermott, D. 1974. Assimilation of new information by a natural language-understanding system. Technical report no. 291, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA.

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence*, pp. 143-160.

McGuire, R., Birnbaum, L., and Flowers, M. 1981. Opportunistic processing in arguments. *Proceedings of the Seventh IJCAI*, Vancouver, B.C., pp. 58-60.

Meehan, J. 1979. *The Metanovel: Writing Stories by Computer.* Garland, New York.

Milne, R. 1982. Predicting garden path sentences. *Cognitive Science*, vol. 6, pp. 349-373.

Minsky, M. 1963. Steps toward artificial intelligence. In E. Feigenbaum and J. Feldman, eds., *Computers and Thought*, McGraw-Hill, New York, pp. 406-450.

Minsky, M. 1968. Introduction to *Semantic Information Processing*. MIT Press, Cambridge, MA, pp. 1-31.

Minsky, M. 1975. A framework for representing knowledge. In P. Winston, ed., *The Psychology of Computer Vision*, McGraw-Hill, New York, pp. 211-277.

Minsky, M. 1979. The society theory of thinking. In P. Winston and R. Brown, eds., *Artificial Intelligence: An MIT Perspective, Vol. 1*, MIT Press, Cambridge, MA, pp. 421-450.

Minsky, M. 1980. K-lines: A theory of memory. *Cognitive Science*, vol. 4, pp. 117-133.

Newell, A. 1982. The knowledge level. *Artificial Intelligence*, vol. 18, pp. 87-127.

Newell, A., and Simon, H. 1963. GPS, a program that simulates human thought. In E. Feigenbaum and J. Feldman, eds.. *Computers and Thought*, McGraw-Hill, New York, pp. 279-293.

Norman, D. 1981. A psychologist views human processing: Human errors and other phenomena suggest processing mechanisms. *Proceedings of the Seventh IJCAI*, Vancouver, B.C., pp. 1097-1101

O'Rorke, P. 1983. Reasons for beliefs in understanding: Applications of non-monotonic dependencies to story processing. *Proceedings of the 1983 AAAI Conference*, Washington, DC, pp. 306-309.

Pereira, F., and Warren, D. 1980. Definite clause grammars for language analysis -- A survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, vol. 13, pp. 231-278.

Pople, H. 1973. On the mechanization of abductive logic. *Proceedings of the Third IJCAI*, Stanford, CA, pp. 147-152.

Quillian, M. 1968. Semantic memory. In M. Minsky, ed., *Semantic Information Processing*, MIT Press, Cambridge, MA, pp. 227-270.

Reichman, R. 1981. Modeling informal debates. *Proceedings of the Seventh IJCAI*,

Vancouver, B.C., pp. 19-24.

Rieger, C. 1975. Conceptual memory and inference. In R. Schank, ed., *Conceptual Information Processing*, North-Holland, Amsterdam, pp. 157-288.

Rieger, C., and Small, S. 1979. Word expert parsing. *Proceedings of the Sixth IJCAI*, Tokyo, pp. 723-728.

Riesbeck, C. 1975. Conceptual analysis. In R. Schank, ed., *Conceptual Information Processing*, North-Holland, Amsterdam, pp. 83-156.

Riesbeck, C., and Martin, C. 1986. Direct memory access parsing. In J. Kolodner and C. Riesbeck, eds., *Experience, Memory, and Reasoning*, Lawrence Erlbaum, Hillsdale, NJ, pp. 209-226.

Riesbeck, C., and Schank, R. 1978. Comprehension by computer: Expectation-based analysis of sentences in context. In W. Levelt and G. Flores d'Arcais, eds., *Studies in the Perception of Language*, John Wiley, Chichester, England, pp. 247-293.

Sacerdoti, E. 1977. *A Structure for Plans and Behavior*. American Elsevier, New York.

Schank, R. 1971. Finding the conceptual content and intention in an utterance in natural language conversation. In *Proceedings of the Second IJCAI*, London, England, pp. 444-454.

Schank, R. 1973. Identification of conceptualizations underlying natural language. In R. Schank and K. Colby, eds., Computer Models of Thought and Language, W. H. Freeman, San Francisco, pp. 187-247.

Schank, R. 1975. *Conceptual Information Processing*. North-Holland, Amsterdam.

Schank, R. 1977. Rules and topics in conversation. *Cognitive Science*, vol. 1, pp. 421-441.

Schank, R. 1982. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, England.

Schank, R., and Abelson, R. 1975. Scripts, plans, and knowledge. *Proceedings of the Fourth IJCAI*, Tbilisi, Georgia, U.S.S.R., pp. 151-157.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

Schank, R., and Birnbaum, L. 1984. Memory, meaning, and syntax. In T. Bever, J. Carroll, and L. Miller, eds., *Talking Minds: The Study of Language in Cognitive Science*, MIT Press, Cambridge, MA, pp. 209-251.

Schank, R., Collins, G., and Hunter, L. In press. Transcending inductive category formation in learning. To appear in *The Behavioral and Brain Sciences*.

Schank, R., Lebowitz, M., and Birnbaum, L. 1980. An integrated understander. *American Journal of Computational Linguistics*, vol. 6, pp. 13-30.

Schank, R., Tesler, L., and Weber, S. 1970. Spinoza II: Conceptual case-based natural language analysis. Research report AIM-109, Stanford University, Dept. of Computer Science, Stanford, CA.

Shwartz, S. 1980. The search for pronominal referents. Research report no. 10, Yale

University, Cognitive Science Program, New Haven, CT.

Slobin, D. 1966. Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior*, vol. 5, pp. 219-227.

Small, S., Cottrell, G., and Shastri, L. 1982. Toward connectionist parsing. *Proceedings of the 1982 AAAI Conference*, Pittsburgh, PA, pp. 247-250.

Sussman, G. 1975. *A Computer Model of Skill Acquisition.* American Elsevier, New York.

Tate, A. 1977. Generating project networks. *Proceedings of the Fifth IJCAI*, Cambridge, MA, pp. 888-893.

Thorne, J., Bratley, P., and Dewar, H. 1968. The syntactic analysis of English by machine. In D. Michie, ed., *Machine Intelligence 3*, American Elsevier, New York, pp. 281-309.

Tyler, L., and Marslen-Wilson, W. 1977. The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, vol. 16, pp. 683-692.

Waltz, D., and Pollack, J. 1984. Phenomenologically plausible parsing. *Proceedings of the 1984 AAAI Conference*, Austin, TX, pp. 335-339.

Weizenbaum, J. 1966. ELIZA. *Communications of the ACM*, vol. 9, pp. 36-45.

Wilensky, R. 1978. Understanding goal-based stories. Research report no. 140, Yale University, Dept. of Computer Science, New Haven, CT.

Wilensky, R. 1983. *Planning and Understanding: A Computational Approach to Human Reasoning*. Addison-Wesley, Reading, MA.

Wilks, Y. 1976. Parsing English II. In E. Charniak and Y. Wilks, eds., *Computational Semantics*, North-Holland, Amsterdam, pp. 155-184.

Winograd, T. 1972. *Understanding Natural Language*. Academic Press, New York.

Winograd, T. 1977. On some contested suppositions of generative linguistics about the scientific study of language. *Cognition*, vol. 5, pp. 151-179.

Woods, W. 1970. Transition network grammars for natural language analysis. *Communications of the ACM*, vol. 13, pp. 591-606.

Woods, W. 1973. An experimental parsing system for transition network grammars. In R. Rustin, ed., *Natural Language Processing*, Algorithmics Press, New York, pp. 111-154.

Zeigarnik, B. 1927. Das Behalten erledigter und unerledigter Handlungen. *Psychologische Forschungen*, vol. 9, pp. 1-85.

END

9-87

DTIC